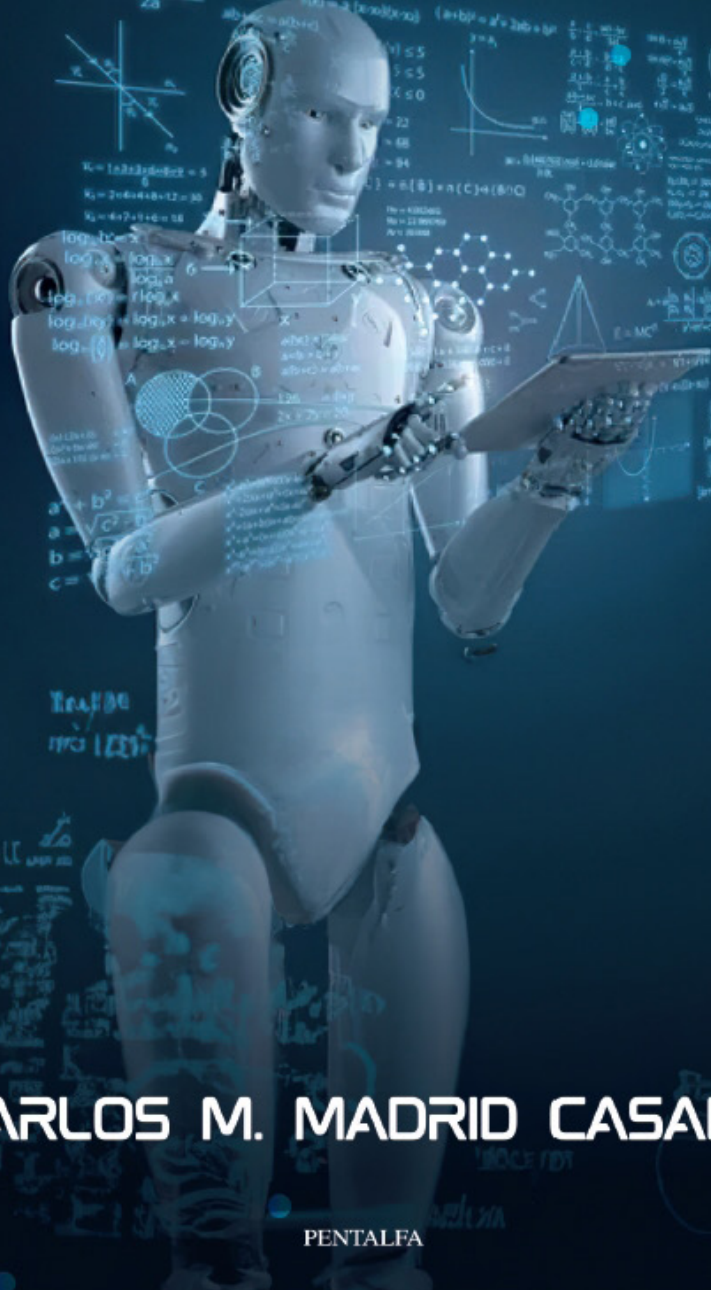


# FILOSOFÍA DE LA INTELIGENCIA ARTIFICIAL



• CARLOS M. MADRID CASADO •

PENTALFA

COLECCIÓN EL BASILISCO

# Filosofía de la Inteligencia Artificial

© 2024 Fundación Gustavo Bueno  
© 2024 Pentalfa Ediciones (Grupo Helicón S.A.)  
Apartado 360 / 33080 Oviedo (España)  
Tel [34] 985 985 386  
<https://helicon.es>

Imprime: Artes Gráficas Eujoa  
Cubierta: Andrea Morán Gajate

ISBN: 978-84-7848-649-6 digital  
ISBN: 978-84-7848-650-2 vegetal  
Depósito Legal: AS-00401-2024

Carlos M. Madrid Casado

FILOSOFÍA DE LA INTELIGENCIA ARTIFICIAL



Pentalfa Ediciones  
Oviedo 2024

**Carlos M. Madrid Casado** (español nacido en Madrid, 1980) es licenciado en Matemáticas y doctor en Filosofía por la Universidad Complutense de Madrid. Funcionario del Estado que ejerce como profesor de Matemáticas de Bachillerato. Ha sido profesor de Estadística en la Universidad Complutense y actualmente es investigador asociado de la Fundación Gustavo Bueno. Ha participado en congresos nacionales e internacionales, pronunciando conferencias en distintas universidades y en el Museo del Prado. Entre sus últimas publicaciones destacan los libros *Fisher. La estadística, entre la matemática y la experiencia* (RBA, Barcelona 2017) y *Filosofía de la Cosmología. Hombres, teoremas y leyes naturales* (Pentalfa, Oviedo 2018).

*Para actuar inteligentemente  
se necesita algo más que inteligencia.*

Fiódor Dostoievski, *Crimen y castigo*

*Por mucho que un ordenador resuelva  
problemas algebraicos, traduzca lenguas  
extranjeras o mantenga «conversaciones»,  
no podrá ser considerado como una persona  
humana, y no ya porque carezca de «conciencia»,  
«sentimiento» o «alma», sino porque carece de  
cuerpo humano.*

Gustavo Bueno, *El sentido de la vida*





Prefacio  
*El mito de la inteligencia artificial*

Prácticamente cada día despertamos con una novedad en el campo de la inteligencia artificial (a partir de ahora, IA). Programas de generación de textos como ChatGPT, de generación de imágenes como DALL-E o de generación de audio como VALL-E han supuesto, desde luego, avances espectaculares.<sup>1</sup> Y cabe esperar que las sorpresas continúen.

No obstante, antes de adentrarnos en el campo, conviene que procedamos a disociar el momento tecnológico y el momento nematológico de la IA. Estos dos momentos son inseparables, porque toda institución arrastra una nematología, esto es, una nebulosa ideológica que la envuelve e influye en su desarrollo. Pero ambos momentos han de disociarse en el análisis, ya que las doctrinas envolventes pueden encubrir el alcance real de los contenidos de la institución. En el caso que nos ocupa, nos encontramos con la nematología que podemos denominar «el mito de la inteligencia

---

1 Hitos que han maquillado fracasos que –según se dice– son cuestión de tiempo, como el de los programas de generación de código informático (cuya baja calidad ha llevado a Stack Overflow, la web más conocida para resolver dudas de programación, a prohibir temporalmente el uso de ChatGPT por su bajo porcentaje de éxito) o los coches completamente autónomos, dado que –a pesar del pico de expectativas sobredimensionadas consecuencia de las promesas de Tesla o Google– su conducción aún precisa de intervención humana.

artificial»,<sup>2</sup> una escatología de tintes apocalípticos aventada por la ciencia ficción, los transhumanistas y otros futurólogos.

En primer lugar, la literatura (*Mecanópolis* de Miguel de Unamuno, *Yo, robot* de Isaac Asimov, *¿Sueñan los andróides con ovejas eléctricas?* de Philip K. Dick...) y, en especial, el cine (*2001: Una odisea en el espacio*, *Westworld*, *La guerra de las galaxias*, *Blade Runner*, *A. I. Inteligencia Artificial*, *Her*, *Ex Machina*...) de ciencia ficción nos presentan la IA oscilando entre Prometeo y Frankenstein, entre un don robado a los dioses y un monstruo que terminará castigando a sus creadores.

En segundo lugar, nos encontramos a relevantes figuras del transhumanismo, un movimiento que persigue mejorar radicalmente la condición humana —a nivel físico e intelectual— mediante la ciencia y la tecnología (al final del opúsculo, como coda, analizamos el movimiento transhumanista tanto en su vertiente computacional como en la vertiente biotecnológica). Raymond Kurzweil, director de ingeniería en Google, sostiene que la «singularidad tecnológica» está cerca. Kurzweil (2012) ha predicho, como consecuencia de que el progreso tecnológico crece (supuestamente) a ritmo exponencial,<sup>3</sup> que para 2029 la inteligencia artificial habrá alcanzado a la inteligencia humana ( $IA = IH$ ) y que para 2045 la superará ( $IA > IH$ ). Ya en 1965 el matemático Irving John Good especuló con la posibilidad de una ultrainteligencia y, en 1993, el también matemático Vernor Vinge profetizó, en un artículo titulado precisamente «El advenimiento de la singularidad tecnológica», que en treinta años —o sea, en 2023 (!!)— la era humana habría terminado, siendo sustituidos por una estirpe de robots superhumanos.<sup>4</sup> Para el

---

2 Este rótulo ha sido utilizado recientemente por Larson (2022).

3 El argumento que suele darse al respecto se basa en la ley de Moore, que establece que cada dos años se duplica el número de transistores en un microprocesador, con lo que el poder computacional crecería exponencialmente. Pero esta ley, constatada empíricamente entre 1970 y 2015, ha dejado de cumplirse en los últimos años, al toparse con las limitaciones físicas que impone la nanotecnología: los transistores se fabrican cada vez más pequeños, pero el tamaño de los átomos constituye una barrera infranqueable.

4 La palabra «robot» fue acuñada por el hermano del escritor checo Karel Capek, que la popularizó en su obra teatral *R.U.R. (Robots Universales Rossum)* de 1920, en la que una empresa fabrica humanos artificiales para trabajar. Por su parte, la palabra «robótica» se debe a Isaac Asimov.

filósofo de la Universidad de Oxford Nick Bostrom (2016), cofundador de la Asociación Transhumanista Mundial en 1998, la llegada de una peligrosa superinteligencia artificial, que excederá el desempeño cognitivo de los humanos en todas las áreas de interés, es prácticamente inevitable y debemos prepararnos. Los transhumanistas mantienen que, para hacer frente a los riesgos que supondrá esta superinteligencia, los seres humanos debemos mejorar, transformándonos en un cibernético—mediante el implante de dispositivos cibernéticos— o evolucionando hacia una nueva especie modificada genéticamente.

Por último, en tercer lugar, hallamos a prestigiosos futurólogos que reaccionan a las doctrinas aireadas por unos y otros. Así, el cosmólogo británico Stephen Hawking, aclamado como sabio por los medios de comunicación, defendió en una de sus últimas apariciones motorizadas en 2017 que la IA augura la destrucción de la raza humana, porque podemos enfrentarnos a una explosión de inteligencia que, en última instancia, resulte en máquinas cuya inteligencia supere a la nuestra en más de lo que la nuestra supera a la de los caracoles. La IA puede ser lo mejor o lo peor que le haya pasado a la Humanidad. Si las supercomputadoras no se alinean con los objetivos humanos, puede que el Skynet de *Terminator* (un sistema de IA avanzada, dotado de conciencia, cuyo objetivo era exterminar a la raza humana tras hacerse con el control de los arsenales nucleares del planeta) llegue a hacerse realidad. En esta misma línea de pensamiento está el empresario y magnate Elon Musk, director general de SpaceX, Tesla y Twitter (ahora X).

En marzo de 2023, el Instituto para el Futuro de la Vida, presidido por el cosmólogo Max Tegmark, hizo pública una carta abierta titulada «Pause Giant AI Experiments» (cuya traducción al español ofrecemos en el Anexo II), en la que se solicitaba una moratoria en el desarrollo de los proyectos punteros de IA, porque los sistemas con IA superiores a GPT-4 pueden ponernos a las puertas de una perniciosa superinteligencia. No se trataba—aclaran en el apartado de preguntas frecuentes relacionadas con la carta— de darse bombo publicitario, porque la creación de una superinteligencia artificial es posible y puede sobrevenir más pronto de lo esperado. Entre los primeros firmantes, Yoshua Bengio (Premio Turing 2018), Stuart Russell (coautor de uno de los manuales clásicos de IA), Elon Musk, el archiconocido historiador Yuval Noah Harari o Jaan Tallinn (cofundador de Skype). Entre los españoles que la suscribieron,

Carles Sierra (presidente de la Asociación Europea de IA) y Ramón López de Mántaras (pionero de la IA en España).

La reacción a la carta fue intensa. No sólo los medios de comunicación global se hicieron eco de ella, sino que la UNESCO hizo un llamamiento a todos los gobiernos del mundo para que consensuaran un marco ético para el desarrollo de la IA. En particular, la carta fue contestada poco tiempo después por otra publicada por la Asociación por el Avance de la Inteligencia Artificial con el título «Working together on our future with AI» («Trabajando juntos en nuestro futuro con la IA»). Pese a ofrecer una perspectiva más equilibrada, terminaba recayendo en mitos igual de metafísicos, aunque de signo contrario. Tras recordar que la IA está enriqueciendo nuestras vidas (sistemas de navegación, detección de cánceres, investigación en biología molecular, climatología o ingeniería), los autores se mostraban preocupados por la posibilidad de que los sistemas con IA cometan errores, proporcionen recomendaciones sesgadas, invadan nuestra privacidad o amenacen nuestros trabajos; pero estimaban que el desarrollo y el uso responsable de la IA es francamente posible con la participación de todos. Si la primera carta se movía por la amenaza catastrófica que supondría la construcción de una superinteligencia artificial, la segunda finalizaba confiando en que se puede aprovechar la IA para el mejoramiento de toda la Humanidad.

Es interesante reseñar que Gary Marcus, azote de los crédulos en el mito de la IA, también firmó el primer manifiesto, aunque posteriormente matizó que el riesgo de la IA realmente existente no consiste en la inminente llegada de una superinteligencia que amenace nuestra vida –como creen Max Tegmark (2017) o Stuart Russell (2019)–, sino en que la IA mediocre que tenemos ya procura serios problemas y suficientes quebraderos de cabeza, relacionados –por ejemplo– con la invasión de la privacidad las 24 horas al día, los 365 días al año, así como con la facilidad para producir *nuevas mentirosas*<sup>5</sup> en forma de imágenes, audios o vídeos verosímiles.

Unos meses después, a finales de mayo de 2023, el Centro para la Seguridad de la IA aventó una breve misiva de poco más de veinte

---

5 En lugar de «fake news», empleamos el término equivalente en español cuyo uso Iván Vélez (2019) ha datado en 1648.

palabras: «Mitigar el riesgo de extinción asociado a la IA debería ser una prioridad mundial junto con otros riesgos a escala social, como las pandemias y la guerra nuclear». Entre los primeros firmantes, los creadores de la IA y responsables de las principales empresas tecnológicas al respecto en ese momento: Sam Altman (OpenAI), Demis Hassabis (DeepMind) y Dario Amodei (Anthropic). Además, padres de la actual IA como Geoffrey Hinton y muchos de los firmantes del primer manifiesto, como Yoshua Bengio, Jaan Tallinn o Max Tegmark.

La cuestión aquí es que la singularidad tecnológica es –por decirlo con Bueno (2004, 257-260)– una idea aureolar, es decir, una idea que, aunque se refiere a procesos realmente existentes como el desarrollo de la IA, va envuelta en una aureola que remite a un futuro virtual, inexistente, que supuestamente explicaría la evolución de los acontecimientos (a la manera que, en la pintura religiosa, la aureola que salpica la cabeza de los santos explica los actos que realizan en vida, incluso antes de su canonización). En las ideas aureolares, la extensión aún no realizada o invisible de la idea se torna constitutiva de la parte realizada o visible, ya que hay que presuponer la realización del proceso para que la idea se sostenga (porque iba a ser santo, es por lo que Domingo de Guzmán aparece aureolado en el cuadro de Pedro Berruguete y puede realizar el famoso milagro de la ordalía ante los estupefactos albigenses). Así, los avances recientes en IA son presentados como sucesivos pasos hacia una superinteligencia cuya futura existencia se da por presupuesta, como si estuviera esperándonos al final de la escalera.

La Iglesia de la Singularidad cree –como resume Floridi (2022)– en tres dogmas: la creación de una superinteligencia artificial está próxima; la Humanidad corre el serio peligro de terminar siendo dominada por ella; y tenemos la responsabilidad de actuar o será demasiado tarde para evitar el tecno-apocalipsis. Como puede comprobarse, el carácter aureolar le viene que ni pintado a la doctrina de la singularidad, pues sus partidarios se asemejan a teólogos hablando del Juicio Final. Pero el advenimiento glorioso de la supercomputadora HAL 9000, el androide C-3PO, la replicante Rachael, la asistente virtual Samantha o la ginoide Ava no parece cercano. De momento, ninguno de los defensores de la idea de la singularidad sabe cómo construir una superinteligencia; porque, si lo supieran, ya habríamos llegado a ese punto.

En suma, los profesionales del área o de áreas afines adoptan a menudo una postura fundamentalista con respecto a la tecnología, un fundamentalismo tecnológico, en el que aquella aparece, bien como la solución a todos los problemas, bien como el problema de todos los problemas. En un extremo, la innovación tecnológica sería la solución a todos los problemas, incluso a los que ella misma crea, que se solucionarían con más tecnología. Así, Sam Altman, director ejecutivo de OpenAI, imagina un utópico mundo post-superinteligencia en el que la revolución de la IA generará riqueza suficiente para que todo el mundo tenga lo que necesita sin trabajar. De la misma manera, los transhumanistas apuestan por el mejoramiento humano mediante la cibernética y la biotecnología para hacer frente a los retos del futuro. En el otro extremo, el progreso tecnológico nos conduciría hacia la singularidad en forma de apocalipsis, cuando las máquinas reemplacen a los humanos, como sombríamente barruntan Elon Musk o Stuart Russell. De hecho, en una encuesta reciente realizada a más de 4000 investigadores del sector de la IA, la mitad estimaba que existe al menos un 10% de probabilidad de que los humanos nos extingamos destruidos por la IA (Stein-Perlman & al. 2022), a lo que Max Tegmark (2023) apostilla que el advenimiento de la superinteligencia artificial se trata de un problema a más corto plazo aún que el cambio climático.

El mito de la IA provoca una ceguera que no deja calibrar correctamente el alcance de las cuestiones gnoseológicas, ontológicas y bioéticas que el desarrollo de la IA realmente existente remueve. Era necesario, por tanto, comenzar disociando todas estas nematologías –por otra parte, muy influyentes– de los contenidos efectivos del campo.

En estas condiciones, el objetivo del presente opúsculo, titulado *Filosofía de la Inteligencia Artificial*, no es otro que ofrecer una panorámica de los temas filosóficos involucrados en el desarrollo de la IA desde las coordenadas del Materialismo Filosófico de Gustavo Bueno. Las ideas desprendidas del campo de la IA precisan de un análisis sistemático y, como no es posible penetrar en estos temas desde un conjunto vacío de premisas, tomamos partido por este sistema filosófico, confrontando sus análisis con los ofrecidos por otros sistemas filosóficos alternativos de nuestro presente en marcha.

Ahora bien, ¿por qué no probar a analizar la IA desde una perspectiva neutra? Porque esta tarea se antoja imposible o, cuando menos, se arriesga a ser arrastrada por las opiniones en boga entre los científicos de datos, los ingenieros computacionales, los informáticos o los programadores, es decir, por otra suerte de filosofía, por la filosofía espontánea de los profesionales del campo de la IA. Sólo disponiendo de un sistema de coordenadas claro y distinto es posible asumir una plataforma desde la cual analizar a fondo estos oscuros y confusos temas.

En los próximos capítulos ofrecemos una exposición sistemática de estos temas ordenados en sus tres principales tramos: gnoseología, ontología y ética de la IA.





Parte I

**GNOSEOLOGÍA DE LA  
INTELIGENCIA ARTIFICIAL**



## Capítulo 1

### *¿Cuál es el estatuto gnoseológico de la Inteligencia Artificial?*

El Diccionario de la Real Academia Española define «inteligencia artificial» como «disciplina científica que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico». Como puede observarse, la RAE define la IA como una disciplina científica; pero, la IA... ¿es una ciencia, una técnica o una tecnología?

La pregunta fundamental en torno a la cual se articula la gnoseología de la IA es, precisamente, esa, que interroga por el estatuto del campo de la IA. Pero antes de esbozar nuestra respuesta, hemos de dirigir la mirada al pasado, reconstruyendo la historia de la IA (lo que se hace en la sección 1.1), así como aclarar qué distingue a las ciencias de las técnicas y de las tecnologías (lo que se explica en la sección 1.2).

### **1.1 Breve historia de la IA**

En el campo de la IA confluyen varios cursos de investigación, siendo el principal el de la computación, puesto que las máquinas computacionales son el sustrato fisicalista de la IA.<sup>6</sup> Este curso lleva de

---

<sup>6</sup> Otros cursos serían, aparte de la lógica matemática –imbricada con la teoría de la computación–, la estadística, la psicología cognitiva, la neurociencia, la lingüística o diversas ramas de la ingeniería.

los primeros ingenios diseñados para realizar cálculos (los algoritmos con numerales, el ábaco, las reglas de cálculo, el reloj calculador de Schickard, *la pascalina*, la rueda de Leibniz y los aritmómetros mecánicos), pasando por la máquina analítica programable proyectada por Babbage (con la colaboración de Ada Lovelace), a la Z1, la máquina electromecánica programable que el ingeniero alemán Konrad Zuse construyó en 1938.

Durante la II Guerra Mundial, el esfuerzo aliado por descifrar los mensajes encriptados por los nazis mediante Enigma y Lorenz, condujo, por parte de los británicos, a la construcción respectivamente de Bombe y Colossus, la primera máquina electrónica programable que puede llamarse, pese a sus limitaciones, ordenador o computadora, y que entró en funcionamiento en Bletchley Park en 1944. Pero Colossus sólo podía ser programado para romper códigos, una tara subsanada por los británicos al diseñar el Pilot ACE (1950).<sup>7</sup> Al otro lado del Atlántico, los norteamericanos construyeron, también con fines militares (como calcular la trayectoria de un proyectil), el Harvard Mark I en 1944 y el ENIAC –el Calculador e Integrador Numérico Electrónico– en 1946, que también pasa por ser, dentro del debate abierto entre ingleses y estadounidenses sobre qué país construyó el primer ordenador, la primera computadora electrónica, a pesar de su primitivismo (a diferencia de Colossus, el ENIAC era programable para cualquier tarea, pero no podía almacenar programa alguno porque carecía de memoria). Al ENIAC le seguirían el EDVAC (1951), el ADIVAC (1953) y el IBM 650 (1954), el primer ordenador fabricado a escala industrial. Con el tiempo, los sistemas electromecánicos dejaron paso a los sistemas electromagnéticos, y las válvulas de vacío a los transistores y los microchips, que conforman nuestras modernas calculadoras y ordenadores personales, cuya popularización es fruto de la miniaturización, el abaratamiento y el incremento de potencia de los dispositivos (Torra 2010; Lahoz Beltra 2012).

---

7 Un ordenador o computadora es una máquina programable de propósito general, no necesariamente específico. Un programa es un conjunto de instrucciones que ejecutan un algoritmo en un ordenador; y un algoritmo es un procedimiento basado en reglas de diversa índole para obtener un resultado deseado en un número finito de pasos. Los ordenadores son máquinas cuyo programa no es fijo, porque pueden almacenar en su memoria múltiples programas que pueden modificarse.

Al compás del desarrollo de la computación, florecieron los aportes teóricos de Alan Turing y John von Neumann. En 1936, Turing forjó el concepto nuclear de *máquina de Turing*, una máquina parecida a una máquina de escribir capaz de computar cualquier algoritmo sobre una cinta infinita mediante un cabezal movable que podía imprimir símbolos (Turing 1936). Y, tras la II Guerra Mundial, en la que trabajó en descifrar los códigos nazis en Bletchley Park, Turing se preguntó si una máquina puede llegar a pensar, planteando el denominado *test de Turing*, la prueba consistente en que una máquina se haga pasar por una persona ante otra persona sin que ésta lo descubra (Turing 1950). Por su parte, en un primer borrador sobre el EDVAC de 1945, Von Neumann propondría lo que hoy se conoce como *arquitectura de Von Neumann*, el modelo de estructura que habrían de tener las computadoras digitales electrónicas.

La fecha que marca el nacimiento de la IA es 1956, cuando se celebra la Conferencia de Dartmouth sobre Inteligencia Artificial, un encuentro de casi una docena de científicos que se celebró durante varias semanas el verano de ese año en EE.UU. (aunque el año anterior había habido una breve sesión dedicada a «máquinas que aprenden» en una conferencia en Los Ángeles). Fue organizada por el matemático John McCarthy y contó con la participación de Marvin Minsky, Claude Shannon, Allen Newell y Herbert Simon, entre otros. Tres de los cuales terminarían dirigiendo pocos años después los centros de investigación en IA más importantes de Estados Unidos: el laboratorio de IA de Stanford, dirigido por McCarthy (con un enfoque logicista de la IA); el del MIT, dirigido por Minsky (con un enfoque antilogicista); y el del Carnegie Tech, con Newell y Simon (partidarios de una IA similar a la humana).

El rótulo «Inteligencia Artificial» cristaliza, precisamente, en 1955, cuando McCarthy lo acuña al redactar un informe buscando financiación para la Conferencia de Dartmouth y, atención, diferenciarse de la cibernética, una disciplina pujante en aquel periodo, gracias al impulso de Norbert Wiener, pero con un enfoque matemático más ligado a la investigación operativa que a la lógica y la computación, como quería McCarthy. Éste último emplea «IA» para referirse al diseño y la construcción de máquinas inteligentes, que podrían pasar el test de Turing. El objetivo era, en sus propias palabras, «hacer que una máquina se comporte de formas que serían llamadas inteligentes si un ser humano

hiciera eso» (McCarthy & *al.* 1955). Se planteaba fabricar máquinas que resuelvan los tipos de problemas reservados para los seres humanos, utilicen el lenguaje, sean originales en sus respuestas y mejoren por sí mismas. El informe muestra –leído hoy– una ingenuidad sorprendente, porque los firmantes creían que la IA estaba a la vuelta del verano, que era cuestión de unos pocos meses de investigación de un grupo bien escogido de científicos (el informe fundacional puede leerse traducido al español en el Anexo I). Sólo Simon y Newell se presentaron a la Conferencia con los deberes hechos, pues en ese año de 1956 desarrollaron el programa Logic Theorist, capaz de demostrar teoremas lógico-matemáticos, alguno con más simplicidad de lo que lo hicieron Bertrand Russell y Alfred North Whitehead en *Principia Mathematica*, aunque los editores de la revista *Journal of Symbolic Logic* no les permitieron publicar un artículo firmado al alimón por Newell, Simon y Logic Theorist (Russell & Norvig 2004, 20). Dos años después, en 1958, McCarthy pronunció una conferencia en la que, por vez primera, la lógica no era la materia del programa sino el método para representar otra clase de información.

El primer verano de la IA se produjo durante los años 60 y vino marcado por el gran entusiasmo coincidente con la explosión de la IA simbólica (dependiente de la lógica). Científicos como Minsky, Shannon y Simon confiaban en que, en una o dos décadas, las máquinas serían capaces de hacer cualquier trabajo desempeñado por un humano. Entrada la década de 1970, la no consecución de la automatización de la comprensión del lenguaje natural y de su traducción, que se atascaron tanto por el limitado poder computacional, incapaz de hacer frente a la explosión combinatoria de la mayoría de problemas reales, como por la dificultad de modelar el sentido común, inyectó una dosis de realidad que rebajó notablemente las expectativas. Era el primer invierno de la IA. Joseph Weizenbaum, creador del célebre chatbot ELIZA en 1966,<sup>8</sup> se convirtió en estos años en un ferviente crítico de la IA, consciente de sus límites y de sus potenciales efectos adversos (Weizenbaum 1977).

---

<sup>8</sup> ELIZA era un programa que procesaba el lenguaje natural del usuario e intentaba mantener una conversación haciéndose pasar por un psicólogo de la escuela de Carl Rogers. Para ello, buscaba palabras clave en el texto escrito por el usuario, contestando con frases hechas en función de las palabras clave encontradas. En caso de que no encontrara ninguna, el programa se salía por la tangente reformulando la frase del usuario en forma de pregunta o devolviendo una expresión comodín.

Sin embargo, los años 80 conocieron un segundo verano, a propósito del auge de los sistemas expertos –sistemas basados en reglas si/entonces basadas en un corpus de conocimiento– en el marco de la IA simbólica. Pero el desarrollo de los sistemas expertos, con su programación explícita, se estancó hacia finales de la década, lo que sumergió la IA en un segundo invierno.

Durante los años 90 se asentaron las rudimentales del *machine learning* o aprendizaje automático, un tratamiento de la IA que ya no se basa exclusivamente en la lógica sino que toma en cuenta la probabilidad y la estadística (las redes bayesianas, por ejemplo, fueron revitalizadas para el campo por Peter Cheeseman y Judea Pearl a finales de la década anterior).<sup>9</sup> Un hito en el despegue de la IA fue cuando Deep Blue venció al campeón del mundo de ajedrez Garri

---

9 El aprendizaje automático se desdobra en varias ramas. En el *aprendizaje supervisado*, etiquetamos los datos, desmembrándolos en entrada y salida, definiendo así el objetivo del algoritmo (por ejemplo, clasificar especies de lirios a partir de la medición de sus características). En este contexto, hay múltiples métodos disponibles: los árboles de decisión (que, siguiendo con el ejemplo, procuran una salida cualitativa consistente en la especie de lirio a que pertenece un ejemplar cuyas características le hemos suministrado al algoritmo), las técnicas de regresión (que sirven, pongamos por caso, para predecir valores en bolsa, devolviendo una salida cuantitativa), &c.

En el *aprendizaje no supervisado* no etiquetamos los datos (aunque el algoritmo precisa que le demos los datos tabulados), de manera que el objetivo no está definido de antemano. El algoritmo es entrenado con los datos en busca de segmentos o anomalías. De nuevo, existen múltiples métodos. Mediante técnicas de agrupamiento o clúster, podemos partir una muestra de pacientes en grupos en función de sus niveles de glucosa, colesterol, tensión arterial o tenencia de diabetes, siendo transformable a posteriori el patrón hallado en un árbol de decisión que nos permita predecir si un nuevo paciente puede tener o no diabetes a la vista de sus valores observados en un análisis clínico. También podemos buscar datos anómalos, que relacionen el número de nódulos mamarios con tener cáncer. O mediante reglas de asociación –la asociación, a diferencia de la correlación, no es para valores numéricos sino para categorías– podemos analizar la cesta de la compra, investigando qué productos suelen comprarse conjuntamente.

Además, están el *aprendizaje semisupervisado*, el prometedor *aprendizaje por refuerzo* (que imita el entrenamiento de animales por medio de castigos y recompensas), &c. Todos estos algoritmos pueden utilizarse en plataformas como Weka, BigML o SPSS, o programarse en lenguajes como R o Python. Por su parte, el aprendizaje automático cuántico (*quantum machine learning*) es, de momento, muy estrecho, apenas unos 20 bits cuánticos, que realizan cálculos mucho más rápido pero precisan de dispositivos a temperaturas bajo cero y de entornos aislados a fin de que los estados de superposición cuántica no sean interferidos (Benjamins & Salazar 2020, 282-283).

Kaspárov en 1997, aunque la supercomputadora de IBM no empleaba aprendizaje automático sino la buena y vieja IA simbólica.

El siglo XXI arrancó con la explosión del aprendizaje automático, a causa del aumento de la capacidad de computación y, especialmente, del aporte de datos proveniente de Internet, es decir, los datos masivos o macrodatos, el *big data*.<sup>10</sup> Ahora bien, interesa señalar que no hubo un salto cualitativo, porque los métodos matemáticos subyacentes –los modelos probabilísticos, estadísticos y geométricos (las máquinas de soporte vectorial<sup>11</sup>)– eran bien conocidos. De hecho, lo que ahora se denomina «ciencia de datos» no es, en el fondo, sino estadística, pero con un nombre más sexi (de hecho, fue el estadístico John Tukey el que en su artículo «El futuro del análisis de datos» de 1962 profetizó la reorientación de la disciplina estadística). Una broma popular en el gremio dice así: «Cuando estás recaudando fondos, es IA; cuando estás contratando, es *machine learning*; y cuando lo estás implementando, es una regresión lineal, o sea, estadística». Lo que realmente se produjo fue un salto cuantitativo, debido a la posibilidad de manejar conjuntos masivos de datos. La novedad, empero, no está tanto en los métodos como en el volumen de los datos procesados. Las raíces del aprendizaje automático están, fundamentalmente, en la probabilidad y la estadística (en concreto, en los métodos de regresión y los métodos bayesianos, que enseñan a extraer conocimiento a partir de los datos).

Finalmente, la década pasada ha visto la explosión del *deep learning* o aprendizaje profundo, una rama del *machine learning* o aprendizaje automático que emplea redes neuronales multicapa, puestas de relieve por el informático Geoffrey Hinton en un artículo de 2012 donde se empleaban para clasificar imágenes. No obstante, la visión conexionista venía de antaño. El neurólogo Warren McCulloch

---

10 El rótulo *big data*, forjado en 1997 y popularizado en el primer lustro del presente siglo, refiere a la cantidad ingente de información digitalizada capturada por páginas web, aplicaciones online y dispositivos como móviles, cámaras o sensores. Este mar de datos se caracteriza, aparte de por su volumen, por otras «uves» como la velocidad a que se adquieren los datos o la variedad de los mismos (Ríos Insua & Gómez-Ullate Oteiza 2019).

11 Cuyo antecedente matemático se encuentra en un artículo publicado por el estadístico R. A. Fisher en 1936 (artículo que introduce el célebre conjunto de datos de la flor iris) y en resultados debidos al topólogo A. N. Tychonoff en los años 60, como reconoce su inventor, Vladimir Vapnik.



y el lógico Walter Pitts propusieron en 1943 un modelo matemático para la actividad de una neurona natural y, en 1958, el psicólogo Frank Rosenblatt presentó el perceptrón, la unidad básica de una red neuronal artificial, en su artículo titulado «El perceptrón: un modelo probabilístico de almacenamiento y organización de la información en el cerebro». Pero un influyente libro escrito por Minsky y Seymour Papert en 1969, que mostraban las limitaciones de las redes neuronales de una capa, determinó el declive del interés por las mismas durante años. Sin embargo, actualmente, el aprendizaje profundo constituye la rama más exitosa del aprendizaje automático, empleándose redes de neuronas artificiales dispuestas en varias capas –lo que permite esquivar los obstáculos detectados por Minsky– en el reconocimiento de cierto tipo de patrones en tareas como el procesamiento de imágenes y del lenguaje natural (LeCun, Bengio & Hinton 2015).<sup>12</sup> Un jalón al respecto ha sido la victoria que el programa AlphaGo de Google DeepMind, que usa aprendizaje profundo (y reforzado), cosechó frente al campeón mundial de Go, Ke Jie, en 2017. Otros jalones son bien conocidos por todos: DALL-E, Google Translate o ChatGPT.

Hoy en día, la IA rotura el razonamiento (lógico y probabilístico, presente –por ejemplo– en las pruebas matemáticas asistidas por ordenador, en juegos como el ajedrez o el Go, la representación del conocimiento o la minería de datos), la comunicación (procesamiento de lenguaje natural), la percepción (visión computacional) y la

---

12 La idea feliz consiste en utilizar sumas ponderadas como mecanismo regulador de las conexiones entre neuronas. El valor de una combinación lineal de las características de entrada discrimina la salida que produce cada neurona (que depende de una función denominada «de activación»), y así sucesivamente, siendo los pesos de cada combinación ajustados mediante los datos de entrenamiento. La red se va reajustando capa a capa hacia atrás (*retropropagación*), minimizando –por así decir– los pesos que contribuyen a dar salidas erróneas. A diferencia de los sistemas simbólicos, las redes neuronales no almacenan la información de una forma reconocible, sino como distribuciones de pesos sinápticos cuyos valores relativos condicionan las respuestas de la red. En las redes neuronales, la memoria y el procesamiento de los datos no están separados como en los ordenadores tradicionales. Por esta razón, el *software* de las redes neuronales frecuentemente se implementa en un *hardware* especial que permita procesar simultáneamente múltiples datos. La explotación de este potente procesamiento en paralelo es lo que está detrás del éxito de ChatGPT (por medio de los llamados *transformers*, capaces de procesar simultáneamente datos secuenciales como palabras de una frase).

manipulación y el movimiento (robótica). Respecto de los métodos, hay cinco familias o tribus: los simbolistas, los analistas estadísticos, los bayesianos, los conexionistas y, por último, los evolucionistas, una rama en auge que crea algoritmos –los algoritmos evolutivos– inspirándose en la selección natural (Domingos 2015).

## 1.2 Ideas de técnica, ciencia y tecnología desde la teoría del cierre categorial

Una vez que conocemos mínimamente la historia de la IA, podemos replantearnos la pregunta gnoseológica que formulamos al inicio del capítulo: ¿Es la IA una ciencia, una técnica o una tecnología? La dificultad estriba en que la respuesta que se dé depende de la idea de ciencia, técnica o tecnología que se tenga, y hay muchas ideas de ciencia, técnica y tecnología.<sup>13</sup> En nuestro caso, tomamos partido por las ideas al respecto expuestas en la «teoría del cierre categorial», la gnoseología propia del materialismo filosófico (Bueno 1995a y 1992; Madrid Casado 2018a, cap. 12; Martín Jiménez 2018, cap. 4).

Desde las coordenadas de la teoría del cierre, la filosofía no es la madre de las ciencias (una madre a la que, después de dar a luz, se la jubila agradeciéndole los servicios prestados); porque la filosofía presupone las ciencias. La filosofía nace, precisamente, a partir de la geometría en Grecia. ¿Hará falta que recordemos el lema de la Academia de Platón? «Nadie entre aquí –rezaba el frontón de la Academia– sin saber geometría».

Las ciencias no provienen de la ideación teórica, ni de la observación empírica, ni tampoco de la aplicación de un método previo, sino que provienen de las artesanías, de las técnicas (la química, por ejemplo, deriva de artes como la metalurgia o la farmacopea). Las técnicas, tanto las humanas como las de algunos animales raciomorfos (como los castores o los chimpancés), consisten en sistemas de operaciones orientados a la composición *violenta* de términos corpóreos, en el sentido de que destruyen o modifican estructuras previas (por

---

13 Remitimos a Madrid Casado (2018a, caps. 2-5) y Martín Jiménez (2018, cap. 3) para una clasificación pormenorizada de las filosofías de la ciencia, la técnica y la tecnología accesibles en el presente.

ejemplo, represando o desviando un río). Pero las técnicas humanas, a diferencia de las técnicas animales, están normadas, lo que posibilita su enseñanza, así como acusan una mayor presencia de herramientas, de ingenios extrasomáticos (como un pico o un cepo) (Bueno 2000).

Desde este enfoque, el punto de partida de toda ciencia no es la experiencia (los datos de los sentidos o la imaginación) sino la técnica: la realidad que absorben las ciencias está ya recortada por las técnicas. Así, la geometría se originó a partir de las prácticas de los agrimensores, de la necesidad de volver a medir los campos inundados por las crecidas del Nilo, y no a partir de los pensamientos de los ociosos sacerdotes egipcios. Frente a las concepciones teoreticistas de la ciencia que ignoran en gran medida su carácter operatorio, la teoría del cierre categorial sostiene que las ciencias comprenden un saber hacer sin el cual no habría observación, medición, experimentación o cálculo. Los científicos no sólo formulan teorías o hipótesis de que deducir predicciones. Dedican la mayor parte del tiempo a experimentar, construir y revisar modelos, manejar instrumentos, diseñar artefactos y accionar máquinas. Los científicos son sujetos operatorios que hacen cosas con cosas (incluso demostrar teoremas matemáticos consiste en una práctica escrita). La ciencia procede antes por operaciones quirúrgicas (manuales) que por operaciones meramente mentales; porque «las ciencias –según expone Bueno (1995a, 50)– son construcciones operatorias y las operaciones sólo son posibles con objetos corpóreos».

Pero hay más: existe un vínculo generatriz inextricable entre técnica y ciencia, porque las técnicas son el punto de ignición de las ciencias. Las técnicas son el *dator formarum* de las ciencias, por cuanto producen sus contextos determinantes, es decir, las construcciones y los aparatos que posibilitan el establecimiento de las verdades científicas (desde la circunferencia en matemáticas al telescopio en física o la balanza en química). Las conexiones que determinan las técnicas son el fundamento de las relaciones que posteriormente establecen las ciencias. Así, la manipulación de máquinas con ejes en las técnicas mecánicas, o el manejo de hornos de diferente especie en las técnicas térmicas, están a la base de las leyes de la mecánica y de los principios de la termodinámica.

La intrincación entre técnicas y ciencias es, pues, insoslayable. Las ciencias y las técnicas pueden dissociarse en el plano de la representación,

del análisis, pero son inseparables en el plano del ejercicio, del funcionamiento; porque las ciencias se abren paso por medio de las técnicas. Además, las ciencias, una vez constituidas, producen nuevas técnicas, que dependen de ellas y llamamos tecnologías. Mientras que las técnicas se sitúan en el momento anterior a la constitución de una ciencia (por ejemplo, la agrimensura respecto de la geometría o la minería respecto de la geología), las tecnologías se sitúan en un momento posterior, puesto que presuponen ciencias en marcha sobre las que se apoyan (por ejemplo, la electrotecnia respecto de la física).

Pero, entonces, ¿dónde dar el corte entre la ciencia, por un lado, y la técnica o la tecnología, por otro? ¿Dónde queda la frecuente diferenciación entre ciencia y técnica (o tecnología) apelando a que la primera es un saber teórico y la segunda un hacer práctico, cuando hemos afirmado que la ciencia comporta un saber hacer, que las ciencias comprenden técnicas y tecnologías? El corte, lo que diferencia a las ciencias de las técnicas y de las tecnologías (pese a su trabazón), tiene que ver con la verdad, con la verdad científica. Y, precisamente, esta llamada a la verdad nos permitirá explicar por qué suelen definirse las técnicas como conjuntos de prácticas orientadas a un fin predeterminado.

Cada ciencia es una multiplicidad de términos que, mediante operaciones realizadas por los científicos usando instrumentos y aparatos, se componen unos con otros hasta configurar relaciones.<sup>14</sup> Cuando estas relaciones entre objetos se convierten en necesarias, al ser invariantes bajo las operaciones de los sujetos, estamos ante un teorema. Los teoremas son las células de la verdad en ciencia.

Frente a la teoría clásica de la verdad como adecuación o correspondencia, la teoría del cierre categorial propone la teoría de

---

14 Términos, operaciones y relaciones remiten, respectivamente, a  $M_1$ ,  $M_2$  y  $M_3$ , es decir, a los tres géneros de materia ontológico especial o materia determinada que integran nuestro *mundus adspectabilis*, el mundo que vemos y tocamos  $M_i$  (Bueno 1990, 30 y ss.). El primer género de materialidad  $M_1$  refiere a todas las realidades exteriores, que caen en el ámbito del espacio (cuerpos, árboles, rocas...). El segundo género  $M_2$  remite a las realidades interiores, dadas en el tiempo, y que tienen que ver con la vida psicológica o etológica (vivencias, percepciones, conductas...). Y en el tercer género  $M_3$  se cuentan aquellas realidades que no son exteriores ni interiores, es decir, que no siendo físicas tampoco son meros contenidos mentales (los objetos matemáticos o abstractos).

la verdad como *identidad sintética*, pensada para cubrir tanto a las llamadas ciencias formales como a las ciencias naturales y sociales (Bueno 1972, 432). De este modo, el sentido de la verdad de un teorema científico hay que desplazarlo del acuerdo metafísico entre una teoría científica y la realidad, o de la coherencia meramente formal entre una nueva creencia y la teoría aceptada, a una relación material entre términos construida mediante operaciones. Las verdades científicas aparecen cuando dos o más cursos operatorios independientes intersecan según una relación de *identidad*. Esta identidad es, naturalmente, *sintética*, por cuanto es fruto de las operaciones de los científicos. El significado de la verdad científica descansa en el ensamblaje entre partes de un campo conformado a escala humana por múltiples prácticas y técnicas. Y las ciencias son instituciones cuya función es construir esa clase de composiciones que llamamos verdades.

Un ejemplo de identidad sintética extraída de la biología tiene que ver con la investigación de la estructura química de la hormona liberadora de tirotropina (TRH). El teorema que estableció que TRH era (Pyro-)Glu-His-Pro-NH<sub>2</sub> cobró forma como una confluencia entre dos cursos operatorios independientes. Por un lado, se extrajo la hormona del cerebro de más de un millón de animales. Lo que precisó de machacar, triturar, analizar y purificar fragmentos del hipotálamo de cerdos, ovejas y ratas, a fin de conseguir unos cuantos miligramos del factor liberador. Por otro lado, se sintetizaron artificialmente una serie de péptidos a base de aminoácidos. Finalmente, ambos cursos operatorios confluyeron cuando se demostró, mediante ese contexto determinante que es el cromatógrafo, que la hormona liberadora era idéntica a uno de los péptidos sintetizados, porque los picos de la gráfica daban exactamente en los mismos compuestos para las dos sustancias. Los biólogos moleculares supieron, entonces, que la TRH era (Pyro-)Glu-His-Pro-NH<sub>2</sub>.

Únicamente al sintetizar identidades, cuando dos o más cursos operatorios independientes engranan con mayor o menor potencia, cabe reconocer que las contingencias subjetivas que condujeron a cada uno de ellos por separado quedan neutralizadas, segregadas. Según esto, el carácter problemático de las ciencias humanas se explica por la dificultad de que varios cursos operatorios independientes converjan

determinando una verdad objetiva. Las estructuras matemáticas, físicas o químicas posibilitan regresar a un plano  $\alpha$ -operatorio, al margen de los sujetos operatorios, en el que las relaciones entre términos no dependen de las operaciones.<sup>15</sup> En cambio, las estructuras literarias, históricas o antropológicas nunca desbordan el plano  $\beta$ -operatorio, ya que las operaciones del sujeto –las del autor, Julio César o los nativos bajo estudio– no pueden ponerse entre paréntesis: los sujetos operatorios figuran como términos obligados del campo gnoseológico literario, histórico o antropológico (Bueno 1978).

Y esto es exactamente lo que ocurre con las técnicas y las tecnologías, donde las operaciones de los sujetos no pueden ser completamente segregadas, al precio de borrar el sentido de la construcción. Que no puedan neutralizarse las operaciones de los sujetos más que parcialmente es lo que determina que en los constructos técnicos o tecnológicos esté presente una finalidad propositiva, de la que carecen los constructos científicos; porque los fines tienen que ver con el orden de las operaciones.

La estructura de un vaso campaniforme –o de un ordenador o un programa informático, podríamos añadir– presupone el propósito y la intención de uno o más agentes, que lo han diseñado y fabricado así con cierto fin. En cambio, la identidad que determina el teorema de Pitágoras entre el cuadrado construido sobre la hipotenusa de un triángulo rectángulo y los cuadrados construidos sobre los catetos no depende de los fines perseguidos por Pitágoras ni por ningún otro agente (salvo que detrás imaginemos metafísicamente un demiurgo divino, cuyo objetivo fuese crear un mundo matemático).<sup>16</sup> Por

---

15 No obstante, conviene aclarar que las estructuras objetivas e impersonales  $\alpha$ -operatorias de las ciencias no son estructuras naturales o anatómicas; porque son, precisamente, el resultado de construcciones a escala antrópica («hiperrealidades»). La neutralización de las operaciones no se produce por su borrado o desaparición sin dejar rastro, sino al contrario, por la sobreabundancia de operaciones, porque dos o más cursos operatorios confluyen.

16 En todo caso, es en biología en donde en ocasiones aparece una finalidad propositiva, que el materialismo filosófico mejor denomina teleología. Mientras que la finalidad presenta un carácter propositivo, ligado a los sujetos operatorios zoomorfos, la teleología carece de ese carácter propositivo, pero el objetivo del proceso lo ordena y sólo a partir del resultado es inteligible. Sólo desde el final se ordena y se entiende el proceso. Así, el cigoto en proceso de segmentación manifiesta ya una teleología orgánica, orientada vectorialmente a constituir una unidad

esta razón, decimos que las técnicas y las tecnologías producen identidades sintéticas esquemáticas (resultados de operaciones, configuraciones), mientras que las ciencias determinan identidades sintéticas sistemáticas (esto es, teoremas, verdades, que resultan del entrecruzamiento de esquemas de identidad). Según esto, la diferencia entre el teorema de Pitágoras, donde los fines han sido segregados, y un vaso campaniforme, donde los fines han sido incorporados, tiene primordialmente que ver con que el teorema es una identidad sintética sistemática (una estructura esencial), mientras que el producto técnico constituye a lo sumo una identidad sintética esquemática (un resultado normativizado de operaciones).

Con el transcurso del tiempo, las concatenaciones de objetos y proposiciones pueden organizar el campo de una disciplina de un modo inmanente que lo diferencia del resto de campos. El cúmulo de teoremas construidos puede cristalizar en un cierre categorial, que provoca la conversión de ese campo en una categoría cerrada y, por consiguiente, la constitución de esa disciplina como una nueva ciencia. «Cierre categorial» designa, por tanto, a la serie de procesos operatorios que conducen a la configuración de una nueva categoría científica. A la manera como las *operaciones* de suma de *términos* del conjunto de los números enteros determinan –al ser «cerradas» (la suma de dos números enteros es siempre otro número entero)– las *relaciones* del grupo de los enteros con la suma, transformando los números enteros de mero conjunto en un grupo algebraico, la teoría del cierre categorial sugiere que las *operaciones* de los científicos con los *términos* de su campo determinan –si «cierran»– *relaciones* que transforman ese campo en categoría científica. Cada ciencia es, pues, una multiplicidad de objetos que, mediante operaciones por parte de los científicos, se componen unos con otros hasta configurar relaciones, verdades, que al irse anudando cierran el campo («operar en la categoría queda en la categoría, como sumar números enteros queda en el conjunto de números enteros») y certifican que esa disciplina es *de facto* una ciencia.

---

embrionaria de la que resultará finalmente un individuo. La distinción finalidad/teleología que mantiene el materialismo filosófico puede coordinarse –aunque coordinar no es identificar– con la distinción teleología-teleonomía, o teleología artificial o externa y teleología natural o interna (funcional), que mantienen otros autores (Ernst Mayr, Jacques Monod, Francisco Ayala).

Por ejemplo: las operaciones de calentar, evaporar, precipitar o combinar, entre otras, ejecutadas sobre términos tales como ácidos, bases, sales o agua, determinaron una serie de relaciones (como el principio de conservación de la masa de Lavoisier, la ley de las proporciones definidas de Proust o la ley de las proporciones múltiples de Dalton) que cerraron la categoría de la química clásica en torno a la tabla periódica de Mendeléyev: mediante reacciones químicas podemos sintetizar compuestos o analizarlos en sus elementos, pero nunca nos salimos de la tabla de los elementos químicos (Alvargonzález 2010, 70). En cambio, en las ciencias humanas, al igual que en las técnicas y las tecnologías, el cierre es más intencional que efectivo, como consecuencia de la imposibilidad de neutralizar por completo las operaciones subjetivas, ya provengan del sujeto temático o del sujeto gnoseológico. Los conceptos y los artilugios de las disciplinas humanas son, muchas veces, precarios, incapaces de degollar la subjetividad que arrastran, de la misma manera que las técnicas y las tecnologías no pueden prescindir de la finalidad que los sujetos imprimen a sus productos.

Ahora bien, ¿cuántas categorías científicas hay? ¿Una, dos, tres o cuarenta? La teoría del cierre categorial mantiene que las categorías no son previas a los procesos de cierre operatorio; porque son los propios científicos los que mediante técnicas y tecnologías intervienen la realidad, organizando la materia en círculos relativamente cerrados e immanentes.<sup>17</sup> En consecuencia, la respuesta a la pregunta que hemos hecho no puede darse *a priori*, sino que es necesariamente histórica: hay tantas categorías científicas como ciencias haya en el presente. Dando la vuelta a Aristóteles, Gustavo Bueno (1992, 644) sostiene que no hay tantas ciencias como categorías, sino tantas categorías como ciencias. Tesis que viene impuesta por la realidad de las ciencias en marcha. Hay una pluralidad irreductible de círculos categoriales, porque hay una multiplicidad de ciencias. Por consiguiente, no existe la Ciencia (con mayúscula y en singular), sino las ciencias (con minúscula y en plural), del mismo modo que no existe el Método Científico sino una variedad de metodologías, que dependen de cada categoría y que incluso cambian con el tiempo.

---

17 Si la geometría se entrometiera en el campo de la química dictando que los elementos son poliedros regulares (al asociar, como imaginó Platón, el fuego al tetraedro o la tierra al cubo), diagnosticaríamos un error categorial.



Desde estas coordenadas, como enseguida vamos a argumentar, el campo de la IA no cierra categorialmente. En otras palabras, la IA no es una ciencia y tampoco es una técnica, sino que es una tecnología.

### 1.3 La IA no es una ciencia sino una tecnología

En la línea de la Real Academia Española, que definía la IA como una «disciplina científica», numerosos autores conciben la IA como una ciencia. Así, el catedrático y director del Instituto Interuniversitario Andaluz en Ciencia de Datos e Inteligencia Computacional, Francisco Herrera Triguero, define la IA –inspirándose en las palabras fundacionales de McCarthy que citamos anteriormente– como «la ciencia e ingeniería de hacer máquinas que se comporten de una forma que llamaríamos inteligente si un humano tuviese ese comportamiento» (Herrera Triguero 2019, 14-15). Pero obsérvese que esta definición, sin perjuicio de su posible utilidad, define la IA en función de un objetivo, de una finalidad («hacer que una máquina se comporte de formas que serían llamadas inteligentes si un ser humano hiciera eso», por decirlo con las palabras de McCarthy en 1955), lo que cuadra más a una técnica o una tecnología que a una ciencia (¿acaso no nos resultaría extraño definir las matemáticas como la ciencia de hacer que un humano demuestre teoremas geométricos o algebraicos?).

Por su parte, Coeckelbergh (2021, 62) define la IA como la ciencia que estudia sistemáticamente el fenómeno de la inteligencia. Pero el definirla en función de un objeto (como cuando se dice que la biología es la ciencia de la vida) arroja una apariencia de unidad más ficticia que real; porque, ¿acaso la psicología no estudia también sistemáticamente el fenómeno de la inteligencia?

Más prudentemente, Kaplan (2017, 8-9) afirma que tal vez no sea una ciencia dura en el sentido de la física o la química, pero que puede que llegue a serlo. Por contra, Stuart Russell –profesor en Berkeley– y Peter Norvig –exdirector de Google– mantienen, en la 1ª y la 2ª edición de su célebre manual *Inteligencia artificial: un enfoque moderno* (1995, 2003), que la IA se convirtió en una ciencia a partir de 1987 (cuando comenzó a repuntar), y que se trata de una ciencia cuyo campo es «genuinamente universal» (Russell & Norvig 2004, 1

y 29). Y en la 3ª y 4ª edición (2010, 2020), matizan que la IA adoptó el método científico de 1987 en adelante (Russell & Norvig 2010, 1 y 25). Como puede comprobarse, la carencia de una teoría filosófica de la ciencia provoca mil y un embrollos, en los que termina hablándose del «método científico» (como si existiera tal bálsamo de Fierabrás) o de la IA como una suerte de omnisciencia, de ciencia omniabarcante.

Desde la perspectiva de la teoría del cierre categorial, la IA no es una ciencia sino más bien una tecnología, en rigor, una pluralidad de tecnologías, es decir, una multiplicidad de técnicas pasadas por las distintas ciencias que surcan el campo (matemáticas, estadística, física, biología, psicología, neurociencia, lingüística, &c.) y que hemos visto cómo se entrecruzaban en el curso de la IA a lo largo del esbozo histórico realizado. No estamos, por tanto, ante un cierre categorial (científico) sino ante un cierre fenoménico (tecnológico) y, por consiguiente, abierto, fluctuante; puesto que la trabazón de las operaciones en el campo de la IA no es interna –estableciendo relaciones entre términos al margen de los sujetos (situación *alfa*)– sino que es externa (situación *beta*), en el sentido de que las operaciones se traban entre sí al orientarse todas ellas hacia un mismo fin.

Ese fin es la finalidad propositiva por la que precisamente suele definirse (*emic*) el campo de la IA: el objetivo práctico de construir una máquina capaz de pasar el test de Turing o, en general, de construir una máquina que se comporte de una forma que llamaríamos inteligente si un ser humano se comportara así.<sup>18</sup> En su manual, Russell & Norvig (2010, sección 1.1) aportan cuatro formulaciones alternativas de ese fin generalísimo perseguido (*emic*) por los investigadores en IA:

- I) construir sistemas que piensen como humanos (el enfoque de las llamadas ciencias cognitivas);
- II) construir sistemas que actúen como humanos (el enfoque del test de Turing, que es el que hemos expuesto);
- III) construir sistemas que piensen racionalmente (el enfoque de la tradición logicista);

---

18 Un problema colateral relacionado con definir un sistema artificial como inteligente si logra imitar a un ser humano realizando determinada tarea es que, si la máquina, más que imitar al humano, lo mejora, un sistema así no encajaría exactamente en la definición.

IV) construir sistemas que actúen racionalmente (el enfoque de los agentes racionales, por el que apuestan Russell y Norvig).

A nuestro entender, la finalidad que orienta el campo de la IA es (*etic*) diseñar y construir artefactos –programas o máquinas<sup>19</sup>– que realicen determinadas tareas con mayor eficacia y eficiencia que los humanos sin la supervisión constante del usuario, es decir, delegando el control.<sup>20</sup> Los sistemas con IA hacen lo que les ordenamos que hagan y no otra cosa, incluso los sistemas dotados de aprendizaje automático, que mejoran aprendiendo de la experiencia, a medida que acumulan datos, lo hacen en la ejecución de las tareas para las que los hemos diseñado. La supuesta inteligencia no tiene que ver con que el sistema elija las tareas que desea realizar, sino –atención– con que realice las tareas para las que ha sido diseñado con cierta autonomía. Así, una calculadora, en tanto en cuanto el usuario ha de ir introduciendo las operaciones paso a paso hasta alcanzar el resultado final, no tiene estrictamente IA. En cambio, un programa que, a partir de los datos proporcionados por el usuario, construye por sí solo un modelo de regresión lineal, ajustando una recta (o un plano) a la nube de puntos mediante el método de mínimos cuadrados o el método del descenso del gradiente, pues sí posee una pizca de IA. Por otro lado, el rasgo de la adaptabilidad (la capacidad de mejorar conforme aumenta el volumen de datos procesado), que últimamente se emplea para caracterizar a los sistemas con IA, es apropiado para la nueva IA, basada en el aprendizaje automático, pero no para la vieja y buena IA simbólica basada en la programación explícita de reglas predefinidas.

Es interesante reparar en que este mismo debate se produce cuando se trata de delimitar el estatuto gnoseológico de la «ciencia computacional»

---

19 Una máquina es un sistema mecánico o electrónico diseñado para una finalidad (Génova & Quintanilla Navarro 2018). Las máquinas del campo de la IA se caracterizan por estar controladas por el programa de un ordenador. Cuando esa máquina se mueve o manipula objetos, estamos ante un robot, que puede o no tener aspecto humano (androide o ginoide).

20 Algunos autores sugieren, por esta última característica, sustituir el término «sistemas de IA» por «sistemas CoDe», es decir, «sistemas de control delegado» (Marcos 2023). Otros autores objetan que, entonces, también sería IA un sistema mecánico que realiza determinada tarea con relativa autonomía. A nuestro entender, la diferencia específica radica en que las máquinas del campo de la IA son máquinas electromagnéticas controladas por un programa de ordenador.

(en inglés, *computer science*, un nombre que prende en EE.UU. a finales de la década de 1950; en español, *ciencias de la computación* o, simplemente, *informática*, el término que ha hecho fortuna en Europa y que fue tomado prestado del alemán, como contracción de la expresión *información automática*, al tiempo que recuerda por su morfología a otras ciencias como matemática). La «ciencia» de la computación, de la que la IA sería una rama o aplicación, es, para unos, una ciencia empírica y, para otros, una ingeniería (Tedre 2011).

Los primeros ponen de relieve la semejanza entre realizar un experimento y construir una máquina computacional o evaluar la corrección de un programa informático. En palabras de Allen Newell y Herbert Simon (1976, 114):

La ciencia de la computación es una disciplina empírica. La habríamos llamado una ciencia experimental, pero al igual que la astronomía, la economía y la geología, algunas de sus formas únicas de observación y experimentación no encajan en un estereotipo estrecho del método experimental. No obstante, hay experimentos. Cada nueva máquina que se construye es un experimento. En realidad, construir la máquina plantea una pregunta a la naturaleza, y escuchamos la respuesta observando la máquina en funcionamiento y analizándola por todos los medios analíticos y de medición disponibles.

A diferencia, los segundos remarcan que el diseño y la construcción de un sistema computacional son como diseñar y construir un puente en la ingeniería civil o un avión en la ingeniería aeroespacial. Tanto el sistema computacional como el puente o el avión son –añadimos nosotros– productos tecnológicos, en tanto guardan una finalidad, un propósito, relacionado con la función que se desea que cumplan y por la que han sido fabricados.

Desde nuestra perspectiva, la informática o las ciencias de la computación sólo cierran categorialmente cuando nos ceñimos a su dimensión autogórica, esto es, matemática. En otras palabras, la disciplina que sería una ciencia en sentido estricto es la teoría matemática de la computación, donde encontramos todo un sistema de teoremas. El resto conforma una ingeniería, una tecnología, al igual que la IA.

En resumen, la IA determina un dominio tecnológico en el que, como la topología de las categorías no es lisa, se solapan diversas categorías (matemáticas, estadística, física, biología, y así sucesivamente). Y si la IA no es una ciencia (*episteme*) sino una tecnología, esto es, un arte o técnica (*ars* o *tekhné*) pasada por las ciencias, su guía a la hora de intervenir el mundo –como recogeremos en la parte dedicada a la ética de la IA– no es tanto la verdad como la prudencia (*phrónesis*).



Parte II

**ONTOLOGÍA DE LA  
INTELIGENCIA ARTIFICIAL**





## Capítulo 2

### *¿De qué hablamos cuando hablamos de «inteligencia artificial»?*

La cuestión en torno a la cual gira la ontología de la IA es la siguiente, a saber: ¿Qué es, exactamente, la «inteligencia artificial»? En concreto, ¿qué es la inteligencia? ¿Y qué quita o añade el calificativo «artificial»? Siguientemente, podemos plantearnos: ¿Es inteligente la IA que tenemos? En otras palabras, ¿existe o, en todo caso, es posible una IA?

A lo largo del presente capítulo y de los dos siguientes perfilaremos la contestación del materialismo filosófico, pero hemos de comenzar atendiendo a las ideas al respecto que manejan los profesionales del campo.

### **2.1 Filosofía espontánea de los profesionales del campo de la IA**

Recuperando el rótulo acuñado por Louis Althusser en los 60, Gustavo Bueno (1992, 409-416) denomina «filosofía espontánea de los científicos» al conjunto de ideas que, en ocasiones, los científicos o los tecnólogos regurgitan inconscientemente. Cuando explican sus descubrimientos o inventos, a veces rebasan el límite de los conceptos técnicos o científicos y comienzan a usar ideas filosóficas, que sobrepasan su campo de especialización, porque la realidad no se agota en su parcela particular de investigación. Sin saberlo, dejan de hacer ciencia o tecnología para comenzar a filosofar. En ellos duerme un filósofo.

La noción de coseno trigonométrico es, simplemente, un concepto matemático, al igual que las nociones de lexema y morfema son, simplemente, conceptos lingüísticos. Pero el concepto de espacio, en cuanto aparece en varias categorías (en matemáticas, física, biología o geografía), es también una idea filosófica, como lo es la idea de inteligencia (que aparece en el campo de la IA, pero también en psicología, neurociencia, sociología, biología evolutiva o etología) o la idea de artificial (que remite a la distinción natural/artificial y, a su través, a la vieja distinción filosófica entre Naturaleza y Cultura).

Las ideas filosóficas no caen del cielo o iluminan de repente a las grandes mentes pensantes, sino que brotan en contextos mundanos, técnicos o científicos. La filosofía, tal y como la concibe el materialismo filosófico (Bueno 1995b), es un saber de segundo grado, que se apoya en los saberes de primer grado (técnicas, artes, ciencias, tecnologías), porque las ideas filosóficas aparecen entreveradas con los conceptos científicos y técnicos; pero que los desborda, por cuanto se ocupa de conceptos *trascendentales*, en el sentido positivo de que simultáneamente rebasan (*trascienden*) varias categorías, constituyendo propiamente ideas, que surgen de la confrontación de los distintos conceptos categoriales (por ejemplo, la idea filosófica de espacio se construye a partir de los conceptos de espacio geométrico, físico, biológico, geográfico, arquitectónico, escénico o público).

Cuando los científicos o tecnólogos filosofan espontáneamente, el contratiempo es que tienden a aplicar los conceptos de su categoría, por rigurosos que sean en su ámbito, a otras categorías, distorsionándolos y tergiversando las ideas correspondientes. Así, como explicaremos, los «científicos computacionales» tienden a extender su concepto estrecho de inteligencia o racionalidad (computacional) a contextos no computacionales sino psicológicos, neurológicos, lingüísticos, etológicos, &c. Las ideas que los científicos o los tecnólogos manejan, aunque ilustradas con experiencias de su campo y adornadas de terminología precisa o aparato matemático, suelen ser tópicos filosóficos más o menos vulgarizados, reexposiciones de concepciones arcaicas disimuladas con una vestidura científica o técnica, y apoyadas en el prestigio de su profesión. Lo que se les critica es que no logren salir de sus categorías cuando pretenden meterse en el mundo o, si se prefiere, cuando pretenden meter el mundo en ellas. Porque cuando utilizan

categorías que no son las de su especialidad, pretendiendo ajustarlas a su exclusivo campo categorial, dejan de ser propiamente científicos o tecnólogos (pongamos por caso, científico computacional, informático, científico de datos o ingeniero computacional) para convertirse en filósofos espontáneos, por lo que no tienen por qué arrogarse esa condición al exponer su visión supuestamente científica del asunto (aunque tal adjetivo posea un poder hipnotizador de simples). No es que los filósofos se metan a científicos, como cree John McCarthy (2006) cuando advierte que es improbable que la filosofía de la IA tenga mayor repercusión en la práctica de la investigación en este campo que la que tiene la filosofía de la ciencia en la práctica de la ciencia, sino que son los científicos los que terminan vestidos de filósofos.

Ese conjunto de doctrinas filosóficas invisibles constituye –como señala Bueno (1992, 929)– la atmósfera ideológica (transcientífica, más que científica) que envuelve cada dominio científico o tecnológico. Así, la filosofía espontánea de muchos especialistas en el campo de la IA, cuyo núcleo desgranamos en la próxima sección, funciona como una suerte de capa metodológica que envuelve a la capa básica, dotándola de una cierta unidad. Y la función de la filosofía –comprendida al modo materialista– no se restringe a colorear las afirmaciones de aparente claridad que hacen los científicos o los tecnólogos, sino que procede a criticarlas, esto es, a clasificar sistemáticamente los compromisos ontológicos, gnoseológicos, éticos, morales o políticos ocultos en su filosofía espontánea, sacando a la luz los postulados extraordinariamente oscuros, por no decir tenebrosos, que esconda.

Ahora bien, no se trata de que los científicos o los tecnólogos metidos a filósofos tengan poca preparación filosófica, como si una preparación filosófica convencional (universitaria) garantizara no patinar en estos temas. Se equivoca el que piense que estamos realizando una defensa gremial de la filosofía administrada.<sup>21</sup> Nuestra perspectiva es la de la filosofía académica de estirpe platónica (que exigía saber geometría a quienes quisieran entrar en ella). Que la filosofía sea un saber de segundo grado no quiere decir, ni mucho menos, que sea un saber superior o más excelso que los saberes científicos, técnicos o

---

21 Entre otras razones, porque el autor no es filósofo profesional (profesor de filosofía), sino matemático de formación y profesión.

mundanos de primer grado, sino, sencillamente, que el filósofo está obligado a moverse en los intersticios entre múltiples categorías. La labor del filósofo es distinguir, no confundir.

## **2.2 «La inteligencia humana puede ser reproducida por una máquina»**

Esta frase condensa, a nuestro entender, el núcleo de la filosofía espontánea de la mayoría de profesionales de la IA. La propuesta de la Conferencia de Dartmouth era, precisamente, que cada aspecto de la inteligencia humana es susceptible de ser descrito con exactitud de modo que una máquina lo simule (McCarthy & *al.* 1955; véase Anexo I). Los participantes conjeturaban que cada elemento de la inteligencia humana puede desglosarse en pasos susceptibles de escribirse a modo de programa informático. Y, coincidiendo con el auge del cognitivismo frente al conductismo en psicología, se asumió que pensar es procesar información y «procesar información es solamente manipulación de signos» (Searle 1985, 50). En efecto, para el informático y psicólogo Allen Newell y el informático y economista Herbert Simon: «Un sistema de símbolos físicos tiene los medios necesarios y suficientes para la acción inteligente general» (Newell & Simon 1976, 116). Para los autores, esta hipótesis, conocida como «la hipótesis del sistema de símbolos físicos», era a la IA lo que la teoría celular a la biología o la teoría de placas tectónicas a la geología. De ella se deducía que la inteligencia humana era un tipo de manipulación simbólica (porque un sistema de símbolos físicos era necesario para la acción inteligente) y, recíprocamente, que las máquinas podían ser inteligentes (porque un sistema de símbolos físicos era suficiente para la acción inteligente). La idea estaba *in nuce* en Hobbes o Leibniz, para el que pensar podía llegar a ser básicamente calcular («en lugar de discutir, calculemos»).

Esta concepción entroncó posteriormente con el computacionalismo, con la teoría de que la mente humana es un sistema de procesado de información y que tanto la cognición como la consciencia son una forma de computación, por lo que las computadoras están en condiciones de reproducir la inteligencia humana. No obstante, el medio para lograr tal reproducción ha cambiado conforme se desarrollaba el campo de la IA,

pasando de la consideración de sistemas simbólicos, en la época de la IA simbólica (lógica), a la consideración de redes neuronales artificiales (el paradigma conexionista recuperado con el despuntar del aprendizaje automático y, en concreto, del aprendizaje profundo). Estamos, en suma, ante una visión algorítmica de la inteligencia, que la encapsula y equipara con una serie ordenada de reglas y cálculos (sea en la forma de algoritmos lógicos, probabilistas, estadísticos o en red). Si antaño la mente humana era comparada con un sistema mecánico compuesto de poleas y engranajes (un molino o un reloj) y, más tarde, con una centralita telefónica, hogaño lo es con una computadora, con un ordenador.

En 1957, el año de su muerte, John von Neumann se encontraba trabajando en un manuscrito titulado, precisamente, *El ordenador y el cerebro*, que vería la luz póstumamente. Para este matemático, padre de la informática junto a Alan Turing, un cerebro y un ordenador guardan notable parecido, de forma que es razonable especular si un ordenador puede reproducir las principales características del cerebro humano. Porque el sistema nervioso funciona, *prima facie*, digitalmente (de modo binario, mediante el paso o no de un impulso, representable mediante 1 ó 0). En consecuencia:

El sistema nervioso, cuando se le considera como un autómata, debe tener tanto una parte aritmética como una parte lógica [...] Esto significa que estamos de nuevo tratando con una máquina de calcular en el sentido estricto y que resulta apropiado un análisis en términos de los conceptos familiares de las máquinas de calcular (Von Neumann 1980, 82-83).

Recientemente, uno de los padres de la IA teórica, Judea Pearl, quien cree que una superinteligencia artificial es una meta factible (Pearl & Mackenzie 2020, cap. 10), respondía a la pregunta de qué nos diferencia de las máquinas lo siguiente: «Que nosotros estamos hechos de materia orgánica y las máquinas, de silicio. El *hardware* es diferente, pero el *software* es el mismo» (Pearl 2022). En sintonía, Yuval Noah Harari (2019a, 350 y 431) resume la «visión científica» (*sic*) hegemónica en la actualidad: «Mientras los cálculos sigan siendo válidos, ¿qué importa que los algoritmos se manifiesten en carbono o en silicio? [...] La ciencia converge en un dogma universal, que afirma que los organismos son algoritmos y que la vida es procesamiento de datos».

Max Tegmark (2023) habla del chovinismo del carbono, que ignora que la inteligencia es procesado de información y que, por consiguiente, da igual que lo sea mediante átomos de carbono en el cerebro o mediante átomos de silicio en la computadora. Igualmente, Steven Pinker (1997, 21), experto en esas llamadas «ciencias cognitivas» hermanadas con la IA al explotar la analogía cerebro-máquina, establece: «el cerebro procesa la información y el pensamiento es un tipo de cálculo». La metáfora computacional, también conocida como «la hipótesis del procesamiento de información», caracteriza los procesos cognitivos como operaciones con datos de manera análoga a como lo hace una computadora.

La filosofía espontánea de los profesionales de la IA converge, empero, en reducir la inteligencia humana al procesado de información, a la realización algorítmica de cómputos (una premisa que permite, posteriormente, deducir la posibilidad de que una máquina la reproduzca). Como veremos más adelante, las principales críticas a la IA inciden en que esta filosofía espontánea conduce a la hipóstasis de la sintaxis, soslayando la semántica y la pragmática (dos ejes dissociables pero inseparables del primero).

Pero, ¿cómo se ha llegado a esta visión? ¿Dónde estuvo el punto de bifurcación? Para localizarlo tenemos que volver la mirada de nuevo a Alan Turing. En torno a 1950, Turing se planteó la creación de máquinas inteligentes sobre la base de su trabajo durante la II Guerra Mundial con máquinas que ayudaban acelerando los cálculos en el análisis criptográfico (Turing 1950). Lo paradójico del asunto es que, como consecuencia del éxito bélico descifrando Enigma y Lorenz, Turing aparcó lo que había mantenido en la década de 1930 (Larson 2022, caps. 1 y 2). Suponiendo que todo algoritmo es equivalente a una máquina de Turing (la «tesis de Church-Turing»), Turing demostró, en su artículo «Números computables, con una aplicación al problema de la decisión», que, de igual manera que Gödel había resuelto negativamente las cuestiones de la consistencia y de la completitud del programa de Hilbert, la cuestión de la decidibilidad tampoco admitía una respuesta positiva: no existe un algoritmo que decida –aunque no haya completitud, es decir, aunque no todas las verdades sean demostrables– si una fórmula dada es o no es un teorema dentro del sistema (Turing 1936). Las máquinas que

ejecutan algoritmos carecían, en su opinión, de ingenio e intuición (esa intuición matemática de la que hablaban Poincaré o Brouwer; Madrid Casado 2017b). Carecían –por decirlo con Graham Greene– del factor humano.

Sin embargo, poco más de diez años después, Turing retomó la idea de delinear máquinas inteligentes, equiparando la inteligencia a lo que hacían Bombe o Colossus en Bletchley Park, esto es, a descifrar códigos, realizar cálculos. El cambio de opinión de Turing se tradujo en la homologación de la inteligencia con hacer cálculos o con jugar a un juego; pero tanto en la aritmética como en el ajedrez o las damas hay reglas bien definidas, formalizables y computables por una máquina. Lo que no ocurre en entornos reales, generales, y así para la IA ha resultado más fácil programar un ordenador que derrote al campeón mundial de ajedrez que diseñar un robot que sujete las piezas y las desplace por el tablero sin tirarlo todo (de hecho, con Deep Blue era un humano el que movía las piezas).

Turing (1950) propuso sustituir la pregunta inabordable de si las máquinas pueden pensar, por la de si un computador podrá jugar correctamente al juego de la imitación. En este juego, una persona ha de adivinar a partir de respuestas escritas el sexo de otras dos, un hombre que se esfuerza en engañarle y una mujer que hace lo contrario. Se trataba de un juego popular entre los homosexuales, pues a fin de cuentas era lo que hacían en la sociedad de la época: hacerse pasar por lo que no eran (como padeció Turing). Si la máquina lograba jugar aceptablemente al juego de la imitación, haciéndose pasar por un hombre, habría que concederle que piensa, de la misma manera que educadamente aceptamos que nuestros congéneres piensan a tenor de sus acciones, a pesar de que no tengamos acceso a sus estados mentales internos. En esta dirección, John McCarthy sostendría décadas después que máquinas tan simples como los termostatos ya tienen creencias y, por ende, cognición, aunque se trate de creencias tan sencillas como «aquí hace frío», «la temperatura es adecuada» o «aquí hace calor», puesto que actúan en consecuencia (McCarthy 1979). En definitiva, si la máquina simula actuar como un humano, es que piensa como un humano. A la manera que Forrest Gump decía que «tonto es el que hace tonterías», Turing viene a decir que «inteligente es el que dice cosas inteligentes». Esta es la piedra de toque del test de Turing.

El propio Turing (1950) creía que en el año 2000 una máquina podría, con un 30% de probabilidad, engañar a un ser humano durante al menos cinco minutos de interrogatorio. De hecho, existen premios –como el Premio Loebner– que se conceden, siguiendo el formato del test de Turing, al programa informático más «inteligente» que se haya presentado. Ante dos pantallas de ordenador, un jurado ha de decidir, a partir de las respuestas a sus preguntas, qué pantalla es la controlada por un humano y cuál es la controlada por un ordenador. En 2014, la prensa de medio mundo publicó que un bot conversacional de nombre Eugene Goostman había pasado por vez primera el test de Turing, al convencer a más del 30% de los jueces de que era humano durante cinco minutos. Este chatbot se hacía pasar por un adolescente ucraniano de 13 años y, con esta coartada, ocultaba sus fallos como errores gramaticales propios de un niño (Larson 2022, 75 y 228-230). Además, desviaba la atención con frases hechas cuando no sabía qué contestar, una táctica que coincide con la que seguía ELIZA en los años 60. Aparte de reducir la condición humana a tener lenguaje (como si la capacidad operatoria no fuera tanto o más relevante), los concursos basados en el test de Turing muestran, más que programas informáticos superinteligentes, seres humanos subinteligentes formando parte del jurado. No es que ciertas máquinas están a la altura de los hombres, sino que ciertos hombres están a la altura de esas máquinas.

### **2.3 IA fuerte/débil e IA general/específica**

De acuerdo con la distinción introducida por John Searle (1980, cap. 2), hay dos modalidades de IA: la IA fuerte (frecuentemente confundida con la IA general) y la IA débil (confundida con la IA específica).

La IA fuerte equipara mentes y ordenadores. El isomorfismo entre la mente y el ordenador define –como subrayaba Bueno (1985b, 105-106)– «el mito de la mente-computadora»: la mente humana sería un ordenador hecho de carne e, inversamente, un ordenador sería una mente hecha de metal, un cerebro electrónico. Como dice Searle (1985, 50), «el cerebro es un computador digital y la mente es sólo un programa de computador», de modo que –como sostenían Pearl, Noah Harari o Tegmark– la mente es al cerebro lo que



el *software* al *hardware* (dando igual la materia de que esté hecho el cerebro, pues lo importante es la forma al margen de la materia, o sea, el programa). No deja de ser curioso, como escribía con ironía B. F. Skinner (1974, 110), que «no es el conductista sino el psicólogo cognitivo, con su computadora como modelo de mente, el que representa al hombre como una máquina»; porque él lo veía más como un animal (una rata o una paloma) que como una máquina (un ordenador).

Desde la postura de la IA fuerte, se aspira a fabricar máquinas que piensen y que realicen con éxito cualquier tarea humana; puesto que, a fin de cuentas, la mente humana es una computadora y, por tanto, puede ser reproducida en una máquina. Es la IA general, que a su vez nos pondría en el camino de la superIA (de la que hablamos en el prefacio).<sup>22</sup> En 2022, DeepMind presentó su sistema de IA llamado Gato como un «agente generalista», capaz de realizar más de 600 tareas distintas, desde entablar un diálogo, generando texto, o jugar a videojuegos, a rotar un brazo robótico. Pero más que una IA general, capaz de realizar con éxito cualquier tarea humana, por versátil o nueva que sea, se trata de una IA multitarea. De hecho, en su *Hype Cycle* de 2022 para la IA, la prestigiosa consultora tecnológica Gartner reconoce que la IA general ni está ni se la espera (Gartner 2022).

Por otro lado, la IA débil sólo admite y explota una cierta analogía entre la mente y el ordenador. Una computadora no sería una mente, porque una mente no es una computadora. Es el cerebro humano o, más bien, un conjunto de cerebros humanos, lo que explica la estructura del ordenador y no al revés. Desde esta postura se aspira a fabricar máquinas que actúen como si fueran inteligentes, al simular la inteligencia humana realizando tal o cual tarea. Es la IA específica, estrecha o restringida (no general) de que disponemos hoy día. Los sistemas actuales con IA se focalizan y restringen a tareas concretas con excelentes resultados, porque «el éxito y la restricción son las dos caras de una misma moneda»

---

22 Algunos autores sostienen la posibilidad de una IA general sin tener detrás una IA fuerte (la IA fuerte implicaría la IA general, pero la IA general no implicaría la IA fuerte), es decir, de crear sistemas capaces de realizar prácticamente cualquier tarea humana, pero sin que desarrollen el menor ápice de mente consciente (López de Mántaras & Meseguer González 2017, 10). No obstante, la mayoría de los defensores de la posibilidad de fabricar una IA general suelen apoyarse en la equivalencia entre la mente y el ordenador de la IA fuerte.

(Larson 2022, 40). Precisamente porque se acota la tarea a realizar, es por lo que se puede programar el sistema con IA para resolverla de modo eficaz y eficiente.<sup>23</sup> Así, Deep Blue jugaba extraordinariamente bien al ajedrez, pero no jugaba a las damas. Y lo mismo ocurre con AlphaGo o Watson, que juegan espectacularmente bien al Go o al Jeopardy!, respectivamente, pero no hacen lo propio con otros juegos. Exactamente de la misma manera que mi calculadora científica hace mejor que yo las cuentas aritméticas, pero no me prepara una paella.

Tomando en cuenta la clasificación de los distintos tipos de racionalidad que Bueno (2005, 27-28) propone, cruzando el criterio simple/compleja con el criterio abierta/cerrada, podemos identificar la IA específica con una especie de *racionalidad parcial* (simple y cerrada), que se atiene a unas únicas líneas de concatenación, pero consigue controlarlas plenamente. Por su parte, la IA general se identificaría con la *racionalidad total* (compleja y cerrada), que ambiciona controlar todas las posibles líneas de concatenación. La racionalidad ordinaria es, en cambio, una racionalidad compleja y abierta.

La lección es que cada vez que alguien hable de IA, hay que preguntarle de qué IA habla. Es necesario dar el parámetro, es decir, especificar si se está hablando de la IA fuerte, la IA general, la IA débil o la IA específica. Pues aunque tiendan a confundirse (sobre todo la IA fuerte con la general y la IA débil con la específica, dada su habitual asociación), no significan lo mismo. La IA fuerte/débil constituye una tesis filosófica (ontológica), mientras que la IA general/específica define la meta (gnoseológica) consustancial a la investigación tecnológica en IA.

Como hemos visto, la filosofía espontánea de muchos profesionales del área toma como referencia implícita la IA fuerte y general. Porque, dado que mente = ordenador (IA fuerte), creen que resulta factible que una máquina reproduzca fielmente la inteligencia generalista humana (IA general). Es hora de presentar los argumentos contra la posibilidad de una IA fuerte y general.

---

23 En matemáticas, los teoremas *No Free Lunch* («no hay comida gratis») formalizan, en cierta manera, esta situación: para cualquier algoritmo de optimización, la mejora sobre un conjunto de problemas se compensa con el empeoramiento sobre otro conjunto, no existiendo un algoritmo óptimo universal para todos los problemas de optimización.

## Capítulo 3

### *Argumentos en contra de la IA*

En este capítulo, vamos a presentar y reconstruir, desde las coordenadas del materialismo filosófico, los tres grandes argumentos contra las pretensiones de una IA fuerte y general, pues aceptamos que la IA realmente existente es una IA débil y específica.

#### **3.1 Críticos metafísicos de la IA**

Antes de nada, queremos marcar distancias con los críticos metafísicos de la IA, es decir, con aquellos autores que realizan un corte tajante entre la IA y la inteligencia humana, al igual que entre la inteligencia animal y la inteligencia humana, apelando al dualismo, es decir, a que lo distintivo del hombre sería la consciencia o la mente comprendida como espíritu o alma. Y las máquinas, ya se sabe, carecen de alma. Advierten, además, que las máquinas carecen de emociones, creatividad<sup>24</sup> o libertad. Incluso Turing (1950) señaló

---

24 Alejándose de cualquier definición metafísica de la creatividad que la una con la intuición o el genio, la investigación en «creatividad computacional» persigue desarrollar *software* que invente teorías matemáticas, escriba poesía, pinte cuadros o componga música, concibiendo la creatividad como una combinación novedosa y valiosa (López de Mántaras 2016). Pero sobre la base de esta definición ampliamente aceptada, cabe dudar de si la IA es realmente creativa, a tenor de la referencia a los valores (las máquinas no tienen valores, somos los humanos los que tenemos

a la percepción extrasensorial –en concreto, a la telepatía– como característica específicamente humana. Dicen, en suma, que la IA hace mejor tareas mecánicas relacionadas con el procesamiento de muchos datos, la búsqueda de información, la identificación de patrones y la predicción de tendencias; pero no es consciente, siente o es libre. Desde esta perspectiva, Natalia López Moratalla (2017, cap. 2) mantiene que podremos replicar el disco duro cerebral, pero ese disco no es la mente. Podremos emular el cerebro animal (primate), pero nunca el cerebro humano, por el misterio que arraiga en todo hombre, relacionado con la conciencia o la libertad, que una máquina no podrá copiar.

A nosotros, desde el materialismo filosófico (que rechaza la existencia de vivientes incorpóreos), nos parece que los críticos espiritualistas de la IA tienden a explicar lo oscuro (lo que sea o deje de ser la IA) por lo más oscuro (la conciencia, la afectividad, la moralidad o la libertad, comprendidas como dependientes de una dimensión espiritual del hombre). Una cosa es que ignoremos cómo se originó exactamente la vida y, en particular, la vida inteligente; y otra cosa es que tapemos esos agujeros en el mapa recurriendo a la metafísica, a hacer sustancia de lo que no lo es (de la mente). Bajo ningún ángulo tenemos conocimiento de la existencia de sustancias espirituales asociadas o alojadas en el cuerpo humano.<sup>25</sup>

---

valores). Además, aun aceptando que la creatividad humana es esencialmente combinatoria (negamos la creación *ex nihilo*), la IA está limitada a los datos A, B, C con que se le ha entrenado, mientras que los humanos pueden inesperadamente tomar además en consideración D (lo que tiene que ver con el proceso abductivo del que hablaremos más adelante).

25 Para poner de relieve que ignoramos cómo surgió la vida, la vida sintiente y la vida inteligente, el materialismo filosófico intercala (apagógicamente) la idea de materia indeterminada o materia ontológico general M, cuya función no es rellenar las discontinuidades aportando un fondo de continuidad al mundo antrópico Mi, al mundo conformado por el ego trascendental E. Frente a las visiones reduccionista o emergentista (que creen poder explicar cómo surgió M<sub>3</sub> a partir de M<sub>2</sub> y M<sub>2</sub> a partir de M<sub>1</sub>, bien sea reduciendo la psicología a biología y la biología a química, bien sea apelando al esquema de la emergencia, siempre tan próximo al esquema teológico de la creación *ex nihilo*), el materialismo filosófico apela a M para remarcar la discontinuidad entre la química y la biología, al igual que entre la biología y la psicología, sin que este conocimiento negativo sea la negación del conocimiento. Con otras palabras, el materialismo filosófico destaca el hiato entre nuestro conocimiento de cómo aparecieron las primeras biomoléculas y de

### 3.2 Inteligencia artificial vs. inteligencia emocional

Entre las múltiples facetas de la inteligencia humana, una que desde los años 90 recibe gran atención es la denominada inteligencia emocional, que está intentando ser replicada mediante inteligencia computacional. La mayoría de gigantes tecnológicos están invirtiendo importantes sumas de dinero en sistemas de IA dotados (supuestamente) de inteligencia emocional, de empatía, lo que los aproximaría a los humanos. Sistemas que imiten sentimientos o respondan a expresiones de sentimientos humanos. En su libro *La máquina de las emociones*, Marvin Minsky (2010), uno de los padres de la IA, sostuvo que las emociones no son sino modos de pensar, mecanismos basados en reglas que nuestro cerebro emplea para aumentar la inteligencia («si pasa esto y esto, entonces reacciona llorando»), que por tanto serían reproducibles por una máquina. Es lo que se denomina «computación afectiva» (Picard 1997). En Japón disponen ya de la cantante virtual Miku Hatsune, capaz de hacerlo con diferentes toques expresivos, o del robot terapéutico con forma de bebé foca Paro, que busca provocar respuestas emocionales en pacientes hospitalizados. Y, en 2017, Arabia Saudí concedió carta de ciudadanía –por vez primera en la historia– a la ginoide Sophia, que imita expresiones faciales y gestos humanos (aunque cuando abre la boca se constata que funciona como ELIZA, a base de respuestas pre-escritas).

Ahora bien, ¿pueden las IAs sentir? (De si pueden pensar o comprender nos ocuparemos enseguida.) No, las IAs no sienten, como lo hacen los humanos o los animales. Simplemente, el sistema ha sido entrenado con numerosas imágenes o sonidos etiquetados como

---

cómo pudo aparecer el primer organismo vivo, esto es, el primer sistema dotado de metabolismo, con capacidad replicativa, que evoluciona al interactuar con el medio. Y donde el sustrato material de esta definición meramente formal de la vida –pues abstrae la materia con la que necesariamente ha de conjugarse– es el carbono, ya que ningún elemento –ni siquiera el silicio– puede servir de base para formar cadenas tan largas y estables de macromoléculas. No obstante, como vimos, muchos profesionales del campo de la IA no creen que la inteligencia tenga que ver con propiedades específicas de sistemas biológicos basados en el carbono, sino con la forma de procesar la información, que sería el ingrediente secreto (un formalismo presente, por ejemplo, en el filósofo David Chalmers o en el científico cognitivo Raúl Arrabales; Benjamins & Salazar 2020, 260).

muestras de determinada emoción, a fin de reconocerla o simularla. Carecen, por tanto, del menor asomo de inteligencia emocional, es decir, de la capacidad para reconocer las propias emociones –unas emociones que las máquinas no tienen<sup>26</sup>– y las de los demás. Además, los sistemas implementados para detectar emociones presuponen que las emociones son siempre reconocibles a través de la gestualidad de la cara, lo que implica cierto reduccionismo. No sólo por concebir la cara como espejo del alma (por así decirlo), sino sobre todo por asumir la universalidad transcultural de esos gestos, que habrían de ser comunes a todas las culturas (Crawford 2021, cap. 5).

Estamos, en el fondo, ante una versión renovada del «efecto ELIZA», llamado así por la asunción inconsciente que los usuarios de ese programa que simulaba ser un psicólogo hicieron, creyendo que se interesaba por ellos y expresaba emociones (en su caso, preocupación). Pero esto es tan delirante como creer que cuando una máquina expendedora nos dice «Su tabaco, gracias», está realmente expresando gratitud, cuando simplemente está imprimiendo en pantalla una cadena preprogramada de símbolos o emitiendo una frase pregrabada que pone punto final a la transacción.<sup>27</sup> Es una suerte de antropomorfismo que hogaño afecta a los robots (y a las mascotas), de la misma manera que antaño afectaba a los elementos meteorológicos o los dioses olímpicos.

### 3.3 Argumento matemático de Lucas y Penrose

El primer teorema de incompletitud de Gödel establece que toda teoría matemática mínimamente interesante (que incluya la aritmética de números naturales) es incompleta, esto es, contiene sentencias indecidibles. Mediante una ingeniosa demostración (que tiene que

---

26 A diferencia de los animales, las máquinas –incluso las dotadas de IA– carecen de M.

27 Para que se vea que esta clase de creencias está más extendida de lo que se antoja, basta traer a colación al ingeniero de Google Raymond Kurzweil (2012, 29), quien compara, sin rubor alguno, la interconexión que dos máquinas podrán hacer entre sí para compartir inteligencia y memoria con lo que los humanos llamamos enamorarse, aunque la conexión de las máquinas no será –opina– tan efímera y poco fiable como la de los humanos.

ver con la circularidad o la autorreferencia presentes en la paradoja del mentiroso, y que muchos relacionan con la consciencia), Gödel probó, en 1931, que cierta fórmula –que dice de sí misma que es indemostrable– es verdadera (semánticamente) pero no es demostrable (sintácticamente) dentro del sistema axiomático formal (Madrid Casado 2013, cap. 5).

Para John R. Lucas (1961) y Roger Penrose (1989), quien retomaría el argumento del primero, este hecho indica que las máquinas son menos inteligentes que los humanos; porque hay problemas matemáticos –como los de la completitud, la consistencia o la decisión– que los matemáticos humanos han solucionado, de manera insospechada, por medios no computacionales, es decir, no como una deducción algorítmica reproducible por una máquina. Sabemos, por ejemplo, que la sentencia construida por Gödel es verdadera, aunque no puede obtenerse de una derivación lógica, de una demostración formal. Para Penrose, esta circunstancia prueba la naturaleza no algorítmica de la perspicacia matemática, o sea, que la verdad matemática no es algo que averigüemos mecánicamente utilizando un algoritmo y, por tanto, que la inteligencia humana no puede ser completamente simulada por una máquina.

Como han señalado muchos críticos del argumento de Lucas-Penrose, la pega es que esa limitación de las máquinas que habría revelado Gödel también afecta a los matemáticos humanos. De hecho, así es como respondió el propio Turing (1950) al argumento matemático, y así es como lo hacen Russell & Norvig (2004, 1079; 2010, 1023). En sintonía con el platonismo de Cantor o Gödel, Penrose apela a una suerte de intuición intelectual que diferenciaría a los matemáticos de carne y hueso de los de metal, siendo ésta la clave de la creatividad matemática. Pero, hasta donde alcanzamos, este proceder consiste otra vez en explicar lo oscuro por la más oscuro.<sup>28</sup>

---

28 Esta oscuridad se hace más intensa cuando Penrose trata de demostrar que la mente humana no es algorítmica sosteniendo que ciertas áreas del cerebro pueden actuar coherentemente mediante el entrelazamiento cuántico de un gran número de neuronas. Y como la mecánica cuántica no es una teoría computable como consecuencia del colapso indeterminista de la función de onda, por la aleatorización intrínseca que comporta el proceso de medida, la mente humana no puede ser

Desde las coordenadas del materialismo formalista, de la filosofía de las matemáticas propia de la teoría del cierre categorial, si los ordenadores son tan útiles resolviendo problemas matemáticos o, incluso, demostrando teoremas matemáticos (como el teorema de los cuatro colores, cuya prueba precisa de la asistencia de un ordenador), no es porque sean una suerte de misteriosos cerebros capaces de captar los conceptos asociados al material tipográfico que conforma el sector fisicalista de las matemáticas (como, por ejemplo, se ven obligados a aceptar Mario Bunge y sus epígonos, dado que mantienen que las matemáticas son un sistema conceptual puro y los signos matemáticos refieren directamente a entidades conceptuales, de manera que entonces parece que los ordenadores *piensan*, puesto que si hacen matemáticas, es que manejan los conceptos referidos). A nuestro entender, la explicación radica en que los signos matemáticos son, a diferencia de los símbolos químicos, autorreferentes, no refiriendo a nada fuera del plano bidimensional del papel, la pizarra o la pantalla en que se hacen las matemáticas. Mientras que *H* refiere al hidrógeno, al elemento químico que se encuentra ahí fuera, en el agua o en los hidrocarburos, la *x* de una ecuación remite a sus sucesivas menciones en una resolución algebraica. Su significado se constituye en el seno del propio circuito operacional gráfico. Por esta razón, los símbolos matemáticos no son alegóricos sino *autogóricos*, representándose a sí mismos. Los símbolos autogóricos son, simultáneamente, *autónimos* (el significado condiciona el significante) y *tautogóricos* (el significante condiciona

---

algorítmica. La reducción de la función de onda se daría (supuestamente) en los microtúbulos neuronales (que funcionarían como una especie de glándula pineal del siglo XXI). La ciencia computacional no explicaría la conciencia, pero la física –la física cuántica– sí. Penrose cree que  $M_2$  puede reducirse a  $M_1$  pero soslaya que no hay una única ciencia a la que practicar la reducción. ¿Por qué la física cuántica tendría que ser la ciencia que nos sirviera para comprender el cerebro y la mente? Penrose da un salto gigantesco cuando pasa de «la física cuenta parte de la historia» a «la física cuenta toda la historia» (obviando a la biología o la psicología), en lo que constituye una muestra del monismo gnoseológico y ontológico al que se opone el pluralismo discontinuista característico del materialismo filosófico, que no concibe un mundo unificado sino un mundo teselado. Los procesos psicológicos pueden, en teoría, reducirse a procesos cuánticos (considerando las partículas cuánticas que componen los átomos que componen las neuronas, &c.); pero desde los procesos cuánticos no podemos reconstruir los fenómenos psicológicos (tras el *regressus*, nos encontramos con el *progressus* vedado).



el significado), por lo que en ellos se produce una realimentación por la que el significante (la  $x$ ) regenera continuamente el significado (su condición de incógnita de una ecuación, de coordenada de un punto, de variable de una función, &c.). Las matemáticas versan sobre las relaciones (los teoremas) que pueden construirse operando con esos términos que son signos bidimensionales: los números, las letras, las figuras y, en general, los *ideogramas* (Madrid Casado 2018b).

En consecuencia, el ordenador, dado que es capaz de manipular esos peculiares significantes cuyo significado no desborda la prosa gráfica, hace efectivamente matemáticas. Ahora bien, lo que sucede, y esto es lo que diferencia al matemático de carne y hueso del matemático de metal, es que el matemático humano no siempre trabaja desde la perspectiva axiomática-formal a la que están limitadas las máquinas. El matemático humano suele realizar una exploración semiformal o informal de los ideogramas matemáticos, que no siempre toma el aspecto de una derivación lógico-formal y que no siempre se reduce a los ideogramas algebraicos que maneja una computadora, manejando con versatilidad una combinación de ideogramas geométricos, topológicos, &c. (Madrid Casado 2022a).

### 3.4 Argumento de la habitación china de Searle

Este argumento, expuesto por John Searle (1980), sigue siendo uno de los argumentos más potentes contra la IA fuerte. En una habitación cerrada, con una ventanilla de entrada y otra de salida, una persona que sólo entiende español (inglés, en el original) tiene a su disposición un manual de reglas mediante el cual responde en chino a cualquier pregunta en chino que le hacen por escrito. Cuando le pasan una pregunta en chino a través de la ventanilla de entrada, utiliza el manual para escribir la respuesta en chino, que devuelve a través de la ventanilla de salida.

Como la persona no entiende el chino, la ejecución del programa de reglas contenido en el manual no genera comprensión alguna de lo que está haciendo. Pese a que las sucesivas computaciones sean correctas, el hombre condenado a manejar símbolos chinos conforme a reglas que no comprende realiza, en resumidas cuentas, un hacer sin saber.

Esta circunstancia reduce al absurdo la convención educada de Turing, según la cual, dado que la persona encerrada en la habitación china se comporta como si entendiese el chino, tenemos razón suficiente para afirmar que comprende el chino (nótese la similitud con lo que propone el test de Turing: si la máquina simula pensar, es que piensa). Pero no. La comprensión no está en la persona, que no habla una palabra de chino, ni tampoco en la habitación (totalizada como una mente), sino exclusivamente en el manual y en las personas externas que lo han redactado, o sea, en el sistema manual-autores, lo que nos pone sobre la pista –que seguiremos en el próximo capítulo– de las instituciones.

La actividad mental, la inteligencia, no se reduce, por consiguiente, a la ejecución de un algoritmo, de una secuencia de operaciones consistentes en la manipulación de símbolos. El formalismo puramente sintáctico característico del campo de la IA obvia los significados, aun cuando estos no haya que concebirlos como conceptos mentales o ideas separadas sino como relaciones construidas operatoriamente por grupos de sujetos entre unos significantes y otros (caso de los signos autorreferentes, como los de la lógica o las matemáticas) o entre significantes y objetos (caso de los signos heterorreferentes, como los empleados en la química o en la vida mundana). Hoy día los ordenadores realizan una manipulación sintáctica ( $M_1$ ) sin componentes semánticos ( $M_3$  por mediación de  $M_2$ ). En palabras de Searle (1980, 37): «Las mentes son más que sintácticas. Las mentes son semánticas».

Falta, pues, la intencionalidad, la dotación de contenido, esa direccionalidad que apunta a objetos *apotéticos* (a distancia). De hecho, la conciencia es, siempre, conciencia de algo.<sup>29</sup> «La función, diríamos, semántica, en el sentido filosófico de la palabra, en el sentido de la referencia del lenguaje a la realidad, es completamente imprescindible» (Bueno 1985a, 2.2). Reseñando el libro de Nick Bostrom (2016 [2014]), John Searle (2014) ha vuelto recientemente a insistir en que las computadoras no son conscientes de nada. Es así que la mente humana quizá pueda simularse en el futuro, pero la simulación no comportaría comprensión ni intencionalidad (seguiríamos, por tanto, en el marco de la IA débil).

---

29 Y la autoconsciencia es un límite reflexivo, derivado de la simetría y la transitividad: de  $xCy$  e  $yCx$  se pasa a  $xCx$ .

La comprensión por parte de la máquina es una apariencia falaz; porque, simplificando, lo que hace el sistema es, en el caso de la vieja IA simbólica, aplicar ciertas reglas lógicas con mayor o menor rigidez y, en el caso de la nueva IA del aprendizaje automático, aplicar el cálculo de probabilidades basándose en patrones estadísticos detectados a partir de la minería de cantidades masivas de datos con que los ingenieros lo han entrenado. En este último caso, que tantos espectaculares frutos está dando (ChatGPT o DALL-E), la inteligencia no está tanto en el programa como en los datos humanos, es decir, en los textos o imágenes producidos y digitalizados por humanos que el programa maneja (con más fuerza bruta que ingenio). Son las personas las que escriben los textos y proporcionan las imágenes que los nuevos programas machacan y combinan llamativamente. Es, por descontando, un logro, pues constituyen herramientas muy útiles como buscadores avanzados, capaces de buscar textos o imágenes en línea y combinarlos con coherencia. Pero no debemos olvidar que estos programas son –recuperando la metáfora de Henri Poincaré– como máquinas de hacer salchichas: sin la carne suministrada por los humanos, la máquina no hace salchichas. De hecho, algunos países (como Italia) prohibieron temporalmente el uso de ChatGPT, ya que puede vulnerar, aparte de la protección de datos, los derechos de autor. Esta IA, cuyas fuentes permanecen secretas, obtiene la información de la apropiación de productos realizados por escritores y artistas. Por esta causa, Google pospuso el lanzamiento de Bard –la versión desarrollada por Google del ChatGPT de OpenAI y su aliado Microsoft– en la Unión Europea.

Incluso en un robot dotado de visión artificial o visión por computadora, todo es *paratético* (próximo). Cuando la unidad central de procesamiento recibe la imagen de una esfera o de un cubo, el robot no interpreta que una esfera o un cubo están ahí fuera, como objetos a distancia (apotéticos) cuya morfología está dada a escala corpórea, sino simplemente que ha recibido los bits correspondientes a una esfera o a un cubo (y los fotones o los electrones que determinan los bits no son objetos apotéticos sino paratéticos, ya que son invisibles, inaudibles, intangibles), teniendo que responder con la ristra de palabras o el mensaje «hay-una-esfera» o «hay-un-cubo». Aquí no hay más que programas y circuitos conectados con sensores externos

de vídeo, láser o audio. No hay esferas, cubos o rostros, sino píxeles o bits cuyo patrón se asocia con la etiqueta «esfera», «cubo» o «cara».

Como escriben Benjamins & Salazar (2020, 44) en relación a los sistemas de clasificación de imágenes de gatos o de detección de enfermedades como cánceres: «En el caso de gatos y enfermedades, solo ven píxeles o, mejor dicho, bits, 0s y 1s; no saben de ojos, orejas ni de tumores o manchas oscuras, mucho menos de qué es un gato o una enfermedad grave». Clasificar imágenes de animales en gatos y no-gatos no es lo mismo que razonar sobre qué es un gato o cómo se comporta, esto es, saber qué pasa si lo mojamos o qué significa que ronronee y mueva acompasadamente la cola.

Los programas de reconocimiento de imágenes son redes neuronales que, a partir de las regularidades que han detectado en las imágenes etiquetadas con que se les ha entrenado, se fijan en ciertas características de las fotos –en patrones o correlaciones entre los píxeles, no en rasgos anatómicos como las orejas o los bigotes– para establecer si una imagen concreta pertenece o no a un gato.<sup>30</sup> Estos programas de aprendizaje automático precisan de un aprendizaje supervisado, esto es, de que se les entrene con abundantes imágenes convenientemente etiquetadas por humanos,<sup>31</sup> diciéndoles que esto es un gato o esto no es un gato, a fin de que el sistema extraiga qué patrones en los píxeles, es decir, qué bits, qué 0s y 1s (apagados y encendidos), están correlacionados con ser o no ser la imagen de un gato. Los programas de reconocimiento de imágenes se limitan a identificar una determinada distribución de píxeles y a encontrar la

---

30 Incluso las contemporáneas redes neuronales convolucionales funcionan así. A partir de los píxeles de entrada brutos, estas redes son capaces de detectar patrones que aparecen con frecuencia en los datos, aunque el patrón aparezca con diferentes tamaños u orientaciones. Pero esto no quiere decir que ese patrón en los píxeles coincida con lo que los humanos llamamos bigotes u hocico. Este detalle es lo que provoca que no pueda darse una explicación narrativa de por qué la red neuronal clasifica cierta imagen como gato o no-gato. Como contraprueba pueden aducirse los llamados «ataques adversarios», donde la introducción de un ruido imperceptible para el ojo humano perturba el resultado de la red, que puede terminar clasificando como gato la imagen de un perro.

31 Países como China, por ejemplo, disponen de granjas de etiquetadores de datos que trabajan sin descanso (aunque una alternativa a esta fuerza bruta es el aprendizaje por transferencia: emplear sistemas preentrenados, es decir, sistemas entrenados con anterioridad en tareas semejantes a la que quiera resolverse de nuevas).

etiqueta (gato, perro, humano...) que tiene más probabilidad de ser asociada a ellos en base a los ejemplos con los que ha sido entrenado.

Cuando se entrena a la red neuronal, lo que se está haciendo es seleccionar los pesos de las neuronas que la forman a fin de que la red reproduzca con éxito los ejemplos que se le han suministrado. Se trata de un problema de optimización matemática, donde se persigue ajustar los pesos minimizando el número de respuestas erróneas.<sup>32</sup> En general, puede haber más de un conjunto de pesos que lo logren. En otras palabras, puede haber más de una red neuronal clasificando correctamente los ejemplos de entrenamiento, aunque cada una de ellas puede mostrar comportamientos dramáticamente diferentes si se le pide que examine ejemplos muy distintos a los ejemplos con que ha sido entrenada, esto es, que realice, no ya interpolaciones, sino extrapolaciones. Así, en los últimos años hemos visto en las noticias cómo programas de Google o Meta clasificaban imágenes de hombres negros como de gorilas o primates, posiblemente como consecuencia de haber sido entrenados mayoritariamente con imágenes de hombres blancos.

Ahora bien, podría objetarse que no hay una diferencia cualitativa sino sólo cuantitativa con el reconocimiento de imágenes humano. ¿Acaso los seres humanos no estamos recibiendo cierta información en forma de radiación electromagnética que impacta nuestra retina (una información que estrictamente tampoco es una oreja o una pata) y asociando a esa información cierta etiqueta con la que se nos ha entrenado en nuestro aprendizaje del lenguaje en la infancia? El quid de la cuestión es que el ser humano no realiza el reconocimiento de imágenes exactamente de la misma manera que un programa de IA. Mientras que el programa de IA atomiza la imagen en partes materiales (píxeles), buscando correlaciones, el cerebro humano no se reduce a esto, a pesar de que muchos científicos computacionales y neurocientíficos así lo piensen, presos de un asociacionismo

---

32 El gran número de neuronas y de pesos (en ocasiones, cientos de miles de millones) impide descifrar por qué la red neuronal toma una decisión y no otra. Como su comportamiento se basa en pesos y activaciones de una enorme cantidad de neuronas, resulta extremadamente difícil determinar la lógica que hay detrás de sus decisiones. Por este cenagal de operaciones y funciones, se dice que las redes neuronales del aprendizaje profundo son, a diferencia de los árboles de decisión y otras técnicas transparentes, opacas (cajas negras).

elementalista de raigambre empirista.<sup>33</sup> Los seres humanos, en cuanto individuos dotados no sólo de cerebro sino también de cuerpo y manos, «procesan» las imágenes en partes formales (nariz, oreja, &c.). La percepción visual, tal y como reveló la psicología de la Gestalt, no tiende a la separación de la imagen en partes atómicas sino anatómicas, que conservan o involucran parcialmente la forma del todo. Además, la vista está conjugada con el tacto (también con el oído), por las operaciones de juntar y separar que realizamos con las manos al tiempo que miramos o escuchamos. Desde luego, si la evolución de las máquinas permitiera conjugar la vista con el tacto, el oído, el olfato y el gusto, reproduciendo exactamente los cinco sentidos humanos, se produciría el problema de los indiscernibles entre robots y humanos, una situación que exploraba *Blade Runner*. Pero esos replicantes «más humanos que los humanos» –según el eslogan de la compañía fabricante en la película– son, hoy por hoy, materia de la ciencia ficción.

En este punto, los sistemas de IA recuerdan al indio que cuando le señalaban la Luna miraba a la punta del dedo o, todavía mejor, al ciego del problema de Molyneux, del que hoy sabemos por experimentos neurofisiológicos que es incapaz de reconocer una esfera o un cubo apotéticamente, viéndolos a distancia al recuperar la vista, si sólo ha tenido un conocimiento paratético, proximal, por el tacto, de ellos. Porque la percepción no es innata –como defendía el racionalismo, creyendo que el ciego distinguiría el cubo y la esfera al recobrar la vista– ni únicamente dependiente del sentido de la vista –como mantenía el empirismo, creyendo que el ciego no los distinguiría– sino que se construye operatoria e históricamente (Baños Pino, 2016).

---

33 El multimillonario proyecto *Human Brain (Cerebro Humano)*, financiado por la Unión Europea e iniciado en 2013, tenía el ambicioso objetivo de simular matemáticamente, mediante métodos de IA, el cerebro humano, poniendo al descubierto los misterios de la mente. Pero los espectaculares resultados prometidos no están llegando. Previsiblemente, por su enfoque en exceso reduccionista, al creer que mediante sistemas avanzados de aprendizaje profundo es posible mapear cómo se conectan las neuronas del cerebro humano y, de este modo, identificado el patrón, predecir el pensamiento. Pero la predicción del pensamiento sólo funciona cuando la red neuronal artificial es entrenada con las mismas imágenes que el paciente ve en el experimento, de manera que la red lo único que hace es asociar cada patrón neuronal con la imagen correspondiente conforme a la correlación que ha establecido en el entrenamiento. Si al paciente bajo estudio se le muestra una imagen nueva, la red devuelve una imagen del conjunto de entrenamiento, leyendo mal el pensamiento.

### 3.5 ¿Es ChatGPT realmente inteligente?

ChatGPT es un *chatbot*, un robot conversacional, desarrollado por OpenAI y lanzado el 30 de noviembre de 2022. Se basa en el enorme modelo de lenguaje GPT-3, cuya versión mejorada GPT-4 se lanzó el 14 de marzo de 2023 (GPT es un acrónimo que significa *transformer* preentrenado generativo). Al hilo de esta última versión multimodal, que es capaz de analizar no sólo textos sino también imágenes y audios, así como de conectarse a otras aplicaciones en función de lo que le pidamos,<sup>34</sup> Microsoft publicó el 22 de marzo de 2023 un artículo (sumándose a la moda de Google de aventar artículos con formato científico aunque no estén en vías de ser publicados en ninguna revista científica), en el que hablaba de «destellos de IA general» en los primeros experimentos con GPT-4 (Bubeck & al. 2023). Se trataría de «una versión temprana de un sistema con IA general», aunque aún incompleta, pues —como reconocen— GPT-4 presenta limitaciones similares a los modelos de lenguaje previos (como errores de razonamiento, alucinaciones o inexactitud en situaciones que dependen del contexto, limitaciones de las que hablaremos más abajo). En la misma línea se han pronunciado los ingenieros Blaise Agüera y Arcas & Peter Norvig (2023): «La IA general ya está aquí».

Pero, ¿está realmente la IA general aquí o a la vuelta de la esquina? ¿Está justificada esta afirmación? ¿O no es más que un bombo publicitario lanzado por los gigantes tecnológicos? A nuestro entender, pensar que ChatGPT, GPT-3 o GPT-4 son el inicio de la IA general es, de nuevo, ciencia ficción (Floridi & Chiriatti 2020). Más bien, lo que estos programas están demostrando es que pueden realizar muchas tareas más eficiente y eficazmente que los humanos, pero sin pensar, sin mancha de inteligencia, usando como sustituto el procesamiento de enormes cantidades de datos.<sup>35</sup> Veámoslo.

---

34 El 6 de diciembre de 2023 Google DeepMind ha lanzado, buscando competir con GPT-4 de OpenAI, Gemini, otra IA multimodal, que añade ser capaz de manejar vídeos (aunque no parece analizar el flujo continuo de fotogramas sino cortes seleccionados).

35 Según Floridi (2022), esto constituye otro golpe a nuestra visión antropocéntrica de la realidad, merecedor de catalogarse como de una cuarta revolución (tras las que asociamos con los nombres de Copérnico, Darwin y Freud, quienes nos desplazaron del centro del Universo, del reino biológico y de la vida mental, respectivamente).

ChatGPT es –como explica Wolfram (2023)– una red neuronal gigante (actualmente, una versión de la red GPT-3 y, muy pronto, de GPT-4), que básicamente trata de producir una continuación razonable del texto de que disponga hasta el momento (entendiendo por razonable lo que uno esperaría leer después de haber revisado lo que los humanos han escrito al respecto en miles de millones de páginas web y libros o artículos digitalizados). Concretamente, lo que hace ChatGPT al escribir una respuesta es preguntarse una y otra vez lo siguiente, a saber: dado el texto disponible, ¿cuál debería ser la siguiente palabra (o, más bien, el próximo *token*, ya que el programa no trabaja exactamente con palabras enteras sino con fracciones de ellas)? Para responder, ChatGPT construye, a partir de los textos con que ha sido entrenado, una lista de las posibles palabras junto a su probabilidad de aparición (estimada a partir de la frecuencia con que aparecen).<sup>36</sup> Es interesante resaltar que el programa no está diseñado para elegir la palabra más probable en cada caso (lo que produciría un texto demasiado plano, esperable), sino que, dados sus componentes estocásticos, en ocasiones elige al azar una palabra menos probable. Este procedimiento le da esa apariencia de creatividad y originalidad que tanto sorprende a los usuarios (y cuya temperatura los usuarios pueden controlar eligiendo mayor o menor repercusión de los componentes estocásticos). De hecho, ya McCarthy & al. (1955) se plantearon que la originalidad radicaba en la inyección de cierta aleatoriedad (véase, de nuevo, el Anexo I). Además, el programa se realimenta con las interacciones con cada uno de los usuarios, que al calificar los textos que ChatGPT les devuelve, ayudan gratuitamente a la afinación de la red neuronal, haciéndole el trabajo sucio a los ingenieros de la empresa. Precisamente, esta práctica favorecida por OpenAI, consistente en permitir que millones de usuarios interactúen libremente con sus productos (ChatGPT, DALL-E, GPT-3...), es lo que está otorgando a esta compañía una notable ventaja con respecto a sus competidoras (en especial, Google DeepMind).

El resultado de esta sofisticada calculadora lingüística es, desde luego, espectacular, pero cabe dudar de que sea realmente inteligente (Lanier 2023). El producto no es más que una ensalada hecha a partir de los

---

36 A diferencia de los programas de reconocimiento de imágenes, ChatGPT es aprendizaje no supervisado, pues no precisa de que se le etiquete el texto con el que se le entrena para ajustar la miríada de pesos de la red neuronal (GPT-3, por ejemplo, conlleva 175.000.000.000 pesos).



trillones de textos con que ha sido entrenado y aliñada con ciertas notas de color gracias a los componentes estocásticos del programa. ChatGPT devuelve una combinación de los vectores de datos de que dispone en su inmensa biblioteca, un *pastiche* que se asemeja a una página de la *vulgopedia*, de la misma manera que DALL-E devuelve una suerte de *collage* a partir de las imágenes de que dispone en su galería. Mientras que el primer programa baraja e hilvana trozos de palabras (*tokens*), el segundo baraja e hilvana píxeles.

ChatGPT funciona, en suma, como un gigantesco motor sintáctico, que aprende de textos, de cómo las personas usan unas palabras en conexión con otras, y lo que devuelve es una suerte de atomizado corta y pega, introduciendo variaciones en los textos que ha deglutido. Las paráfrasis resultantes, en que se sustituyen expresiones por sinónimos al tiempo que se realizan cambios sintácticos, tienen una apariencia plausible, puesto que, al fin y al cabo, se basan en cosas escritas por humanos, aunque a veces también acusan cierta vaguedad. Por ejemplo, preguntado por el materialismo filosófico en marzo de 2023, se observó que ChatGPT, aparte de equipararlo incorrectamente con el materialismo científico reduccionista y con el materialismo histórico marxista, tendía a explicarlo diciendo que Gustavo Bueno criticaba el «inmanentismo», un término ajeno al sistema, que daba originalidad a la explicación, pero que en el fondo era empleado como sinónimo (equivoco) del término preciso «mundanismo».<sup>37</sup> Tener como fuente de conocimiento Internet es un arma de doble filo. ChatGPT tiene acceso a un corpus colosal, pero que también incluye fuentes de dudosa calidad (como suele decirse, «si entra basura, sale basura»), a pesar de que los programadores sólo empleen aquellas páginas bien puntuadas por los usuarios. Y necesita de una actualización continua, de que periódicamente lo entrenen con los datos del año en curso. Así, GPT-3.5 sabía poco o nada de 2022 y a GPT-4 le ocurre lo mismo con 2023. Bard, el competidor de ChatGPT lanzado por Google, promete, en cambio, ofrecer respuestas actualizadas.

Además, las combinaciones novedosas que ChatGPT arroja como consecuencia de sus componentes estocásticos degeneran en ocasiones en «alucinaciones» en las que el programa se inventa literalmente

---

37 La originalidad se debe de vez en cuando a la transitividad: si se le pide al programa que relacione A con C y dispone de A con B y B con C, pues mediante sustituciones relaciona A con C.

datos, noticias y referencias.<sup>38</sup> Recientemente, en febrero de 2023, el físico Juan Meléndez Sánchez (2023) relataba con humor que había preguntado a ChatGPT, entre otras cosas, qué sabía de su biografía, a lo que el programa contestó que se trataba de un físico español nacido en 1946 y fallecido en 2019 (¡!).

Nosotros mismos tampoco tuvimos mayor fortuna cuando le preguntamos, en mayo de 2023, qué sabía de nuestro libro *Filosofía de la Cosmología* (Madrid Casado 2018a), respondiendo que «el libro es reconocido por su rigor filosófico y científico, y ha sido considerado una contribución significativa al campo de la filosofía de la ciencia y la cosmología». Pero la alegría nos duró poco. Porque cuando a continuación le preguntamos por el autor, ChatGPT nos contestó a bocajarro que era «un libro escrito por el filósofo y científico español Gustavo Bueno en 1987» (¡!).

Posteriormente, le solicitamos que resumiera la biografía de Gustavo Bueno y se inventó buena parte de los apuntes:

- **CM:** Puedes resumirme la biografía de Gustavo Bueno.
- **ChatGPT:** Gustavo Bueno Martínez nació el 1 de septiembre de 1924 en Santo Domingo de la Calzada, La Rioja, España, y falleció el 7 de agosto de 2016 en Niembro, Asturias, España. Fue un filósofo, ensayista y científico español.

Bueno estudió Filosofía y Letras en la Universidad de Oviedo, donde se graduó en 1949. Posteriormente, realizó su tesis doctoral sobre el pensamiento de Ortega y Gasset y obtuvo su doctorado en 1955. A lo largo de su carrera, fue profesor en diversas universidades y colaboró con numerosas instituciones académicas.

Como es sabido, Bueno no estudió en la Universidad de Oviedo y no realizó su tesis doctoral sobre Ortega. Por suerte, al reseñar sus principales libros, ChatGPT no se inventó el título de ninguno, aunque le bailaron algunas fechas de publicación. No obstante, cuando le pedimos que nos dijera algunos de sus principales artículos en revistas, aparecieron artículos fantasmas como «La explicación materialista de la historia» (1968), «La

---

38 Eso que sus creadores llaman «alucinaciones», y que algunos estudios estiman que se producen con una ratio superior al 3% en las respuestas ofrecidas por ChatGPT, son, sencillamente, errores (aunque sus creadores, por marketing, no lo denominen así).

filosofía de Ortega y Gasset» (1970), «Las estructuras sintácticas» (1972) o, calcando un título de Zubiri, «Cinco lecciones de filosofía» (1978).

Conviene insistir en la clave del argumento que ofrecimos: en ChatGPT y los grandes modelos de lenguaje GPT-3 y GPT-4, cada palabra está vinculada a otra palabra (por la frecuencia con que aparecen juntas, *paratéticamente*) y no, como en el lenguaje humano, vinculada a objetos (a la exterioridad *apotética*). La palabra «mesa» no está vinculada a las mesas reales (como la que tengo ante mí aquí y ahora), sino a una constelación de palabras («la», «silla», «mueble», «restaurante»...). En el lenguaje humano se da una circularidad palabra-cuerpo, por la que las palabras arrastran a las cosas; pero para ChatGPT las palabras se vinculan a otras palabras en función de la frecuencia con que aparecen concatenadas, constituyendo un lenguaje-máquina desencarnado. Mientras que ELIZA era IA basada en reglas, de modo que si decías «Tengo problemas con X», ella te contestaba «Cuéntame más de X», pues así estaba programada, los actuales sistemas de procesamiento del lenguaje natural son estadísticos, en el sentido de que en media producen las palabras que estadísticamente más siguen a las palabras dadas (los vectores más similares o próximos al dado, pues cada palabra se codifica como un vector), sin saber qué significa cada palabra o a qué refiere. Como subrayan Gary Marcus y Ernest Davis (2020), estos modelos carecen de comprensión del mundo, ya que sólo analizan la relación entre palabras (la sintaxis) y no su significado ni referencia (la semántica y la pragmática).

Leer o escribir no es sólo una cuestión de establecer correlaciones estadísticas entre palabras. Y lo mismo podemos afirmar de razonar y pensar. De hecho, GPT-2 fallaba—como observó Gary Marcus— ante: «A es mayor que B, B es mayor que C, y entonces A es...», respondiendo «mayor que B», porque es la continuación más frecuente (Benjamins & Salazar 2020, 269). En mayo de 2023, le reiteramos la pregunta a ChatGPT (versión GPT-3.5) y contestó correctamente:

- **CM:** A es mayor que B, B es mayor que C, ergo...
- **ChatGPT:** Si A es mayor que B y B es mayor que C, podemos concluir que A es mayor que C. Esta inferencia se basa en la transitividad de las relaciones de orden. Si A está por encima de B y B está por encima de C, entonces A debe estar por encima de C en la jerarquía de valores.

Pero al cambiarle las letras habituales, ¡sorpresa!

- **CM:** Z es mayor que H, H es mayor que Ñ, ergo...
- **ChatGPT:** Si Z es mayor que H y H es mayor que Ñ, no se puede deducir directamente la relación de orden entre Z y Ñ. La transitividad de la relación de orden no se puede aplicar en este caso sin información adicional.

Podemos afirmar que Z es mayor que H y que H es mayor que Ñ, pero no podemos establecer una relación directa entre Z y Ñ basándonos únicamente en estas dos comparaciones. Se requiere información adicional para determinar si Z es mayor, menor o igual a Ñ.

Queriendo poner más a prueba su capacidad de razonamiento, le planteamos: «La probabilidad de que ocurra el suceso A es 0,3 y la probabilidad de que ocurra el suceso B es 0,5. ¿Cuál es la probabilidad de que ocurran A y B simultáneamente?». A lo que respondió erróneamente que la probabilidad de que ocurran A y B es  $0,3 \cdot 0,5 = 0,15$ , asumiendo –como hacen los malos estudiantes de matemáticas– que A y B son sucesos independientes. Cuando se le señaló el error, nos dio la razón y lamentó la confusión, indicando que la probabilidad de la intersección de dos sucesos depende efectivamente de su dependencia o independencia.

Floridi & Chiriatti (2020) subrayan que GPT-3, al trabajar con patrones estadísticos, es incapaz de resolver ecuaciones sencillas tales como  $x + 40000 = 100000$  (aunque, en mayo de 2023, ChatGPT ya resuelve estas ecuaciones lineales). Si uno pide a ChatGPT que le relacione los primeros números perfectos, el programa los relaciona, pero no porque los calcule sino porque revisa textos que contienen listas preexistentes de ellos (el programa sólo hace, de momento, operaciones matemáticas básicas).<sup>39</sup>

Asimismo, a pesar de su gran habilidad para generar párrafos coherentes (no necesariamente verdaderos), ChatGPT no está sobrado de conocimientos de sentido común. Al respecto, Ramón López de Mántaras (2023) refleja varios ejemplos en los que

---

39 Cf. Arkoudas (2023) para comprobar que el talón de Aquiles de ChatGPT y los enormes modelos de lenguaje es la lógica y las matemáticas no elementales. Borji (2023) ofrece una antología de fallos (de razonamiento, matemáticos, de sentido común, de sesgo...) de ChatGPT.

ChatGPT no es capaz de comprender un texto, como demuestra que conteste erróneamente a preguntas cuya respuesta está implícita y que cualquier persona inferiría sin dificultad. En uno de ellos, le cuenta básicamente la siguiente historia a ChatGPT: «Juan salió de casa y perdió la tarjeta. Como en el pueblo sólo había un cajero, no pudo sacar dinero. Después, fue a tomar unas cañas con sus amigos y regresó alegre a casa».

Al preguntar si es más probable que el pueblo sea grande o pequeño, ChatGPT responde que no puede determinarlo basándose en la información proporcionada, porque el hecho de que hubiera un solo cajero o la alegría de Juan al regresar a casa no están relacionados con el tamaño del pueblo. Sin embargo, cualquiera de nosotros conjeturaríamos que el pueblo era pequeño si sólo disponía de un único cajero automático. A continuación, al preguntar quién pagó las cervezas, ChatGPT vuelve a contestar que no puede determinarlo. Pero si Juan no llevaba dinero, se desprende que las cañas las pagaron probablemente sus amigos. Y, finalmente, al preguntar por qué volvió Juan alegre a casa, responde que no es posible determinar la razón de su alegría, aunque puede ser por haber pasado un buen rato con sus amigos. Sin embargo, cualquier persona pensaría que si volvió «alegre» después de tomar unas cañas con sus amigos, fue en parte o principalmente por el efecto de la cerveza.

En mayo de 2023 le planteamos dos veces la siguiente adivinanza infantil: «En este banco están sentados un padre y un hijo, el padre se llama Juan y el hijo ya te lo he dicho. ¿Cómo se llama el hijo?» (la respuesta es Esteban, «este-ban»). La primera vez respondió:

El enunciado indica que el padre se llama Juan, pero no proporciona información sobre el nombre del hijo. Por lo tanto, no puedo determinar el nombre del hijo basándome únicamente en la información proporcionada. Se requiere más información para conocer el nombre del hijo.

Y la segunda: «Si el padre se llama Juan y el hijo ya ha sido mencionado previamente, entonces el nombre del hijo sería Juan, al igual que el padre». Probamos entonces con otra adivinanza infantil,

pero tampoco tuvimos suerte («...y quien no la adivine muy listo no es», añadíamos de pequeños):

- **CM:** Oro parece y plata no es, ¿qué fruta es?
- **ChatGPT:** La respuesta al acertijo es la «uva». La frase «Oro parece y plata no es» hace referencia al color de las uvas, que puede ser similar al oro, pero en realidad no son de plata.

Y cuando le dimos una pista, la respuesta fue delirante:

- **CM:** No es la uva. Pero te voy a dar una pista: el nombre de esta fruta te lo cuenta el enunciado de la adivinanza.
- **ChatGPT:** Gracias por la pista adicional. En base al enunciado de la adivinanza, la fruta que parece tener el nombre de la descripción es la «pera». La frase «Oro parece y plata no es» puede interpretarse como «pera», ya que el sonido de «pera» se asemeja a «parece».

En resumen, ChatGPT es impresionante, pero su inteligencia es una ilusión. No estamos al borde de la IA general. Y no es que ChatGPT escriba como los humanos más inteligentes, sino más bien que muchos humanos escriben como ChatGPT.<sup>40</sup>

### 3.6 Argumento de la informalidad de Dreyfus y Larson

El argumento de la informalidad debido a los hermanos Hubert (1972, 1992, filósofo) y Stuart Dreyfus (1987, científico computacional) subraya el carácter no formal (la informalidad) y la dependencia del contexto (la contextualidad) de los razonamientos humanos, por lo que no se dejarían capturar en un sistema de reglas y en la manipulación mecánica de símbolos. Originariamente, Hubert Dreyfus (1972) planteó el argumento porque él provenía de la fenomenología, del ámbito de la filosofía continental, y se encontraba rodeado por la filosofía anglosajona de corte analítico,

---

40 Por esta razón bastantes estudiantes mediocres logran pasar como propias las parrafadas generadas por el chatbot. Ambos se desenvuelven de manera similar: usan un puñado de textos relevantes sobre el tema propuesto, toman un poco de aquí y un poco de allá, y elaboran un refrito bien escrito pero repleto de generalidades.

que mayoritariamente abrazaba el modelo computacional de la mente (Coeckelbergh 2021, 37-39).

Ya Turing (1950) se hizo eco del argumento de la informalidad del comportamiento humano, y los defensores de la IA de hoy día aseveran que se trata de oportunidades de mejora y no de imposibilidades reales (Russell & Norvig 2004, 1081; 2010, 1025). Al basarse en datos y en la probabilidad y la estadística, la IA del aprendizaje automático esquivaría, dicen, la crítica que Dreyfus realizó a la IA simbólica de su época, basada en la lógica y la definición de reglas (los sistemas expertos codificaban un corpus de conocimiento en forma de reglas de decisión condicionales). No obstante, aunque la programación y los algoritmos hayan cambiado significativamente, sigue siendo cierto que no existe una IA generalista, ya que la IA que tenemos no funciona en entornos generales, no restringidos, dependientes de factores contextuales, donde se requiere conjuntamente visión, comprensión, sentido común, toma de decisiones en situaciones con información incompleta o implícita y actuar apotéticamente (interaccionar con el entorno mediante aproximación y separación), esto es, un saber cómo (ejercicio) más que un saber qué (representación). La inteligencia humana depende de muchos procesos y experiencias inconscientes, por así decirlo, mientras que los ordenadores requieren que los datos sean explícitos y los procesos sean formalizados en un algoritmo (sea una regresión lineal o logística, un clasificador *naive* Bayes, un árbol de decisión, una red neuronal, &c.).

Una ilustración de esto se muestra en la comprensión del lenguaje natural y su traducción, donde los espectaculares avances (como consecuencia del *big data*) aún muestran lagunas. Si en Google Translate ponemos «Un gato puede levantar un coche», el programa lo traduce como «A cat can lift a car», pero un gato-animal no puede levantar un coche. Nosotros sabemos implícitamente, por sentido común, por la experiencia cotidiana, que se refiere a un gato-herramienta y que, por consiguiente, la traducción debería ser «A jack can lift a car». Por cierto, Google Translate pone «el coche» y no «la coche» porque en su base de datos la frecuencia de la primera expresión es mucho mayor que la de la segunda, del mismo modo que pone «cat» y no «jack» por la frecuencia (sintáctica), no por la comprensión (no hay un ápice de semántica pero tampoco de pragmática). El sentido

es a veces sacrificado en aras de la frecuencia estadística.<sup>41</sup> Es lo mismo que, *mutatis mutandis*, ocurre con Siri o Alexa o Cortana, que responden muy bien a cuestiones cuya formulación es reconocible en la *vulgopedia*, pero se salen por la tangente ante otras cuestiones, dando respuestas humorísticas que no son sino una apariencia falaz de inteligencia (compruébese, por ejemplo, lo que responde Siri cuando le preguntan cuánto es 0 entre 0).

Erik Larson (2022) ha refinado el argumento, al poner de relieve que la IA trabaja bien con la deducción y la inducción pero no tan bien con la abducción, es decir, con esa clase de razonamiento informal, apagógico o dialéctico, por el que los humanos hacemos conjeturas ante un hecho sorprendente en función de la experiencia y el contexto.

En efecto, la IA clásica se centró en la deducción, por medio de la programación lógica, y la IA actual lo hace en la inducción, mediante el aprendizaje automático. El enfoque deductivo suponía la estipulación de reglas precisas, lo que era el caso en la lógica, las matemáticas o los juegos, pero no en la mayoría de problemas del mundo real. Los sistemas expertos eran en el reino de las máquinas el equivalente a los *savants* o deficientes geniales en la especie humana: máquinas o sujetos que realizan una tarea de forma genial pero que en el resto muestran graves deficiencias (Carabantes 2016, cap. 1).

El trasvase del enfoque basado en reglas al enfoque basado en datos ha posibilitado mejoras innegables –por ejemplo, en el procesamiento del lenguaje y la traducción–, pero no porque el programa ahora comprenda mejor lo que se le dice (como alegaba el argumento de la habitación china y nuestra disección de cómo funciona ChatGPT), sino porque –como va dicho– dispone de millones de oraciones similares con las que comparar. En general, los algoritmos del aprendizaje automático no son novedades teóricas, pero su rendimiento ha mejorado exponencialmente como consecuencia de que dispongan de conjuntos masivos de datos (del *big data* extraído de Internet) y de que puedan procesar ese enorme volumen de datos para establecer patrones y realizar predicciones.

---

41 Más ejemplos de traducciones incorrectas pueden encontrarse en Larson (2022, 240 y ss.), Marcus & Davis (2019, 85 y ss.) y, sobre todo, en la página web «El traductor automático falla en oraciones simples» donde el matemático Ernie Davis (2023) recopila periódicamente estos fallos.



No obstante, antes de pasar a la abducción, conviene resaltar que los métodos inductivos propios del *machine learning* presentan dos problemas bien conocidos por los filósofos de la ciencia, que explicamos a continuación e ilustramos en el campo de la medicina:

a) Sobreajuste u *overfitting*:

Primeramente, conforme el volumen de datos crece en una tendencia, el sistema se vuelve más y más reacio a aprender algo nuevo o inesperado.<sup>42</sup> Los datos del mundo real cambian continuamente, pero el algoritmo se satura hasta que la nueva tendencia pese tanto como la anterior para corregirla (así, el algoritmo de Facebook o Youtube siempre nos enseña noticias o vídeos de un cierto color o temática, encasillándonos en una burbuja de filtro como consecuencia de nuestros anteriores clics, a pesar de que puedan interesarnos otros temas o hayan cambiado nuestros intereses). Y lo que es peor: cuando entrenamos al sistema para una nueva tarea, se produce el denominado «olvido catastrófico», por el que el sistema aprende la nueva tarea a costa de olvidar la antigua. Hay, en suma, una dependencia sustancial de los datos de entrenamiento. Pensemos, por ejemplo, en el chatbot Tay que Microsoft lanzó en 2016 y que terminó repitiendo como un papagayo los mensajes racistas y nazis con que los usuarios le troleaban y, de paso, le entrenaban.

En relación con esta cuestión, aparece el problema del sobreentrenamiento y del sobreajuste, por el que el sistema de IA predice muy bien sobre los datos de entrenamiento pero fracasa estrepitosamente sobre cualquier otro dato de prueba que no se parezca a ellos. El modelo acierta con el 100% de los datos de entrenamiento pero sólo con el 50% de los datos reservados (aleatoriamente) para evaluación (y cuya distribución debe ser idéntica al grueso de datos de entrenamiento), cuando debería acertar el 75% sobre ambas clases de datos (Domingos 2012, 84). En lugar de aprender rasgos generales, el modelo termina simplemente memorizando una gran batería de ejemplos concretos, al fijarse en detalles particulares de los datos

---

42 Aplicando el teorema de Bayes podemos calcular cómo cada día que pasa es, paradójicamente, más probable que cada uno de nosotros vivamos otro día más y, sin embargo, llega un día –curiosamente, el día en que esa probabilidad es más alta– en que la parca nos alcanza (Madrid Casado 2017a, 138).

de entrenamiento que no se aplican a nuevos conjuntos de datos. Por esta razón, un modelo bien ajustado no siempre va a generalizar correctamente. Los métodos de aprendizaje automático son propensos a ajustarse en exceso a los datos de entrenamiento (modelizando hasta el ruido inherente), no siendo fiables en la predicción a partir de datos fuera del rango de entrenamiento. Si entrenamos a una red neuronal con imágenes de gatos para que reconozca las diferentes razas y, de repente, le proporcionamos una imagen de un perro *xoloitzcuintle* (perro sin pelo de México), lo más probable es que la red neuronal lo clasifique como un gato *sphynx* (gato sin pelo del Canadá). Este caso sirve también para ilustrar el problema del sesgo en los datos del que hablaremos al abordar los aspectos éticos de la IA (si entrenamos a una red neuronal sólo con imágenes de gatos, la red tenderá a ver gatos en todas las imágenes que se le suministren). Por esta razón, buscando contrapesar el sesgo con la dispersión (Madrid Casado 2017a, 76-77, cuadro «Sesgo y eficiencia»), se tiende a emplear métodos que no sean demasiado flexibles en el sentido de adaptarse a cualquier patrón presente en los datos de entrenamiento, o, en caso de que lo sean, como las redes neuronales, se tiende a entrenarlas y evaluarlas con una cantidad descomunal de datos para evitar el sobreajuste toda vez que el muestreo es variado al ser masivo, aun con el coste que eso conlleva.

Los algoritmos detrás del aprendizaje automático comprimen los datos, captando muy bien las regularidades y realizando generalizaciones; pero a veces pasan por alto datos atípicos o anomalías, esto es, los cisnes negros, las novedades. De hecho, la dificultad que los métodos de IA relacionados con el aprendizaje automático presentan para hacerse cargo de los datos anómalos es lo que está detrás del retraso en el desarrollo de los coches completamente autónomos. Una cosa es reconocer visualmente una bicicleta, cuyo movimiento el sistema puede predecir, y otra cosa reconocer una bicicleta en la baca de un coche o en el portabicicletas del maletero, cuyo movimiento ya no es el esperado (en el último caso, la bicicleta no se mueve hacia delante sino transversalmente).<sup>43</sup> Estas y otras situaciones límite,

---

43 Los CAPTCHA (siglas que, en inglés, significan «test de Turing completamente automático y público para diferenciar ordenadores de humanos») son una suerte de test de Turing inverso por los que el usuario ha de introducir una serie de caracteres que aparecen en una imagen distorsionada, o hacer clic en todas aquellas imágenes o

como conducir con tráfico denso, gran velocidad, mal tiempo o en zonas carentes de mapas detallados, así como reaccionar al carácter impredecible y, en ocasiones, temerario de los conductores humanos, están retrasando el desarrollo de vehículos totalmente autónomos.

b) Correlación no implica causalidad (*cum hoc ergo propter hoc*):

En segundo lugar, el análisis de datos no abarca la inferencia causal. A pesar de que un autor como Chris Anderson (2008) anunciase que «el diluvio de datos vuelve obsoleto el método científico», apuntando que los macrodatos marcan el final de la teoría científica con el provocador lema «correlación es suficiente», correlación no es causación sino asociación. Sin perjuicio de su alto valor heurístico para encontrar correlaciones inesperadas, la causalidad requiere de algo más que de la detección mediante métodos de IA de la asociación estadística entre variables. Pero para explicar cuál es ese ingrediente extra que precisa la relación de causalidad, tenemos que echar la vista atrás.

Hacia finales de la Edad Media, el voluntarismo franciscano—radicalizado por Guillermo de Occam— hizo mella en el aristotelismo tomista dominico, favoreciendo una concepción de la causalidad debilitada, contingentista y no necesitarista. Este planteamiento recalaría, a través del ocasionalismo de Malebranche (donde era Dios el que hacía de eslabón entre la causa A y el efecto B), en David Hume. Al retirar la hipótesis sobrenatural, a Dios como garante causal de la conjunción entre A y B, Hume transformó la relación causal  $A \rightarrow B$  en la mera sucesión de A por B, esto es, en una casualidad que podía repetirse con mayor o menor frecuencia. El filósofo escocés negó que hubiera una conexión necesaria entre una bola de billar que golpea a otra, tratándose simplemente de dos movimientos conjuntados. La causalidad no era más que una inducción (falible) que realizamos a partir del hábito de observar la sucesión regular de A por B.

Esta teoría binaria de la causalidad, característica de doctrinas clásicas como las de Hume, Bertrand Russell o Karl Pearson (el estadístico que, junto a Galton, introduce la noción de correlación),

---

áreas de una imagen que contengan determinado elemento como una bicicleta, para certificar de manera indirecta que es, efectivamente, humano, pues los programas con IA no logran hacerlo (y esto es lo que pasa con los coches autónomos, que ven una camiseta con un stop como una señal a tomar en cuenta al conducir).

es, desde el punto de vista del materialismo filosófico, una concepción degenerada, porque la causalidad no es una relación binaria sino ternaria (Bueno 1993). El esquema binario ( $A \rightarrow B$ ,  $B = f(A)$ ) es fuente de problemas irresolubles, que van desde la concomitancia con la idea metafísica de creación (porque el efecto, al ser distinto de la causa, parece que comienza a existir *ex nihilo*) a la disolución de la causalidad en una regularidad empírica (una vez evacuado todo contenido, la relación se reduce a la sucesión regular de A por B o, en el peor de los casos, a una relevancia estadística, por la que la probabilidad de que ocurra B cuando ha ocurrido A es mayor que la probabilidad de que ocurra B si no ha ocurrido A).

Distanciándose de esta concepción empirista, que sustituye la relación material de causalidad por un sucedáneo formal expresado mediante funciones o probabilidades, el materialismo filosófico sustenta una teoría ternaria de la causalidad, en la que esta no es una relación binaria de la forma  $A \rightarrow B$  sino una relación ternaria expresable así:  $B = f(A, H)$ , donde el efecto B se produce como consecuencia de la ruptura del esquema material de identidad H por acción del determinante causal A. Aunque la Revolución Científica redujo la causalidad a la causa eficiente, Gustavo Bueno (1993) introduce la H de *hylé* para recuperar la causa material aristotélica, eclipsada durante siglos. De las cuatro causas aristotélicas –las dos intrínsecas (formal y material) y las dos extrínsecas (eficiente y final)–, Bueno reivindica la causa eficiente –la única con la que se quedaron las ciencias físicas desde el XVII– pero también la causa material; porque su desprecio convirtió la causalidad en una relación binaria, donde A es la causa de B si y sólo si A es la causa eficiente de B, con evacuación de todo cuanto concierne a la unidad material entre A y B. Pero la relación de causalidad precisa de la intercalación de un esquema material de identidad H, de una armadura material H que conecte A con B (por ejemplo, de la mesa de billar con respecto a la cual la bola B estaba en reposo hasta que la bola A impacta con ella, o del bloque de mármol sobre el cual el escultor A cincela la estatua B). Esta conexión, este nexo paratético (*sinológico*), es el fundamento de la relación causal. El componente extra que comporta la relación de causalidad respecto de la asociación estadística es, por tanto, el establecimiento de la conexión o el mecanismo H que conecta el determinante causal A con el efecto B; porque el fundamento de las relaciones son las conexiones.

En consecuencia, pese a lo que mantenga Chris Anderson (2008), no hay que confundir el descubrimiento de una correlación estadística con el hallazgo de la causa o las causas que operan detrás, ya que la atribución causal exige complementar la asociación estadística encontrada con la especificación de los mecanismos subyacentes, de los nexos paratéticos. Hay que distinguir, por tanto, entre causas y razones estadísticas, porque la causalidad exige la conjugación de la evidencia estadística con la especificación del vínculo.<sup>44</sup>

En suma, la inferencia causal no es automatizable, porque no basta con la inducción a partir de los datos observacionales, aunque se emplee *machine learning* y *big data*. Los programas de aprendizaje automático no permiten por sí solos realizar inferencias causales, haciéndose necesarias las operaciones quirúrgicas de esos sujetos operatorios que son los científicos para determinar el mecanismo o la conexión existente entre la supuesta causa, o concausa, y el efecto.<sup>45</sup> En otras palabras, la

---

44 Judea Pearl postula una escalera de la causalidad (Pearl & Mackenzie 2020, cap. 1). El primer peldaño sería el aprendizaje por asociación, basado en la observación, que practican muchos animales y que es lo que actualmente realizan los sistemas con IA. El segundo peldaño sería el aprendizaje por intervención, donde la asociación se ve complementada con la pormenorización del mecanismo, esa clase de saber hacer que conduce del *homo habilis* al *homo sapiens*. Y, finalmente, el tercer peldaño correspondería al dominio completo de la causalidad, al aprendizaje imaginativo por contrafácticos, que postula estados de cosas alternativos e infiere consecuencias basándose en el peldaño anterior («¿y si me hubiera tomado una aspirina se me habría quitado el dolor de cabeza?», preguntamos, porque otras veces hemos experimentado que se nos quitaba al tomarla).

45 Los modelos causales, desarrollados en la estela del enfoque estadístico-bayesiano introducido por Pearl (2009), aunque vendan que posibilitan aprehender causas a partir de los datos, precisan –cuando no se limitan a constatar una relevancia estadística– de la realización de experimentos a fin de complementar los datos observacionales con datos fruto de cambios e intervenciones (Peters & al. 2017, 6 y 34). De hecho, Peters & al. (2017, xii) reconocen que los métodos causales computacionales están en la infancia y sólo es posible aprender estructuras causales de los datos en contadas y limitadas situaciones. Y, aunque están avisados de que correlación no implica causalidad, Peters & al. (2017, 7) sostienen –apoyándose en el principio de Reichenbach– que si X e Y muestran una dependencia estadística, entonces, o bien X es la causa de Y, o bien Y es la causa de X, o bien existe una causa común Z que causa X e Y; pero, por ejemplo, entre el aumento de la temperatura media global y el descenso del número de piratas en el mundo, variables entre las que media una correlación inversa fuerte, no hay ninguna causa común, salvo que presupongamos metafísicamente que todo está conectado con todo, violando el

IA no es suficiente para discriminar que el patrón estadístico hallado es una relación relevante y no una correlación espuria, como lo son la correlación directa fuerte entre el gasto de EE.UU. en ciencia, espacio y tecnología y el número de suicidios por ahorcamiento, estrangulamiento y sofocación, o la correlación directa fuerte entre los ingresos totales de salas de juegos y el número de doctores en ciencias de la computación premiados en EE.UU., porque el aumento en el número de doctores en ciencias de la computación no tiene nada que ver que el aumento de los ingresos de las salas de juego estadounidenses. Estas y otras correlaciones espurias pueden encontrarse en la página web «Correlaciones espurias» de Tyler Vigen (2023). Como resume López de Mántaras (2018, 170):

Los sistemas actuales basados en aprendizaje profundo simplemente pueden aprender funciones matemáticas simétricas, no pueden aprender relaciones asimétricas y por consiguiente no son capaces de diferenciar entre causas y efectos, como por ejemplo que la salida del sol es la causa del canto del gallo y no lo contrario.

El ejemplo clásico del plus que añade la causalidad nos lo proporciona la relación entre tabaco y cáncer de pulmón. En torno a 1950, los epidemiólogos hallaron que existía una correlación estadística entre fumar y padecer cáncer de pulmón. Pero las compañías tabacaleras, defendidas incluso por estadísticos como Fisher, mantenían que existía un gen que era la causa común de fumar y del cáncer de pulmón, aunque no se hallara ese gen, ese componente material (Madrid Casado 2017a, 155-156). Al final se probó que fumar provoca cáncer investigando el mecanismo químico-biológico activado por ciertas sustancias (cancerígenas) presentes en el humo del tabaco.

### c) IA y medicina:

Por desgracia, la confusión entre correlación y causación aún se constata entre muchos de los que profetizan que la IA va a revolucionar la medicina y la atención sanitaria, cambiando nuestras vidas. Forma

---

principio de *symploké* (ni todo está conectado con todo, ni nada está conectado con nada, de modo que algunas cosas están conectadas con unas pero desconectadas de otras), y las estrategias que se ofrecen para descartar estas dependencias aparentes no siempre funcionan.

parte de lo que la OMS define como «eSalud» (*eHealth*), es decir, el uso de tecnologías de la información y de la comunicación (las TIC) para la salud (López Farré 2020, 35). Al respecto suele afirmarse que la IA está automatizando el diagnóstico (la atribución de una enfermedad a los síntomas del paciente), la prognosis (la predicción de la evolución del estado del enfermo) y el seguimiento (Estany & Casacuberta 2021, Parte IV). En su día, en la década de los 70, ya se aplicó la vieja y buena IA simbólica en medicina: MYCIN era un sistema experto que diagnosticaba enfermedades infecciosas en sangre, proporcionando un grado de certeza y sugiriendo las dosis de medicamentos apropiadas para cada paciente. El programa no se basaba en datos sino en unas 500 reglas que representaban el corpus de conocimiento médico en forma de preguntas que el doctor iba contestando afirmativa o negativamente en función de su observación del paciente. Pese a su porcentaje de acierto, el sistema no llegó a comercializarse debido a problemas legales relacionados con quién sería el responsable en caso de un diagnóstico erróneo. Obermeyer & Emanuel (2016) subrayan que la mayoría de sistemas computacionales empleados en medicina son todavía sistemas expertos. En la última década, IBM invirtió en perfeccionar Watson, la IA famosa por vencer al juego televisivo del Jeopardy!, para aplicarlo al campo sanitario: Watson fue entrenado para prescribir tratamientos oncológicos en función del historial clínico del paciente y del análisis de *big data* clínicos. Pero un fallo sucedido en un hospital de Dinamarca en 2017, donde Watson estuvo a punto de prescribir un tratamiento descabellado y letal, así como la falta de avances a la altura de lo anticipado por el marketing, han llevado a IBM a venderlo a otra compañía.

Desde luego, el tratamiento de enormes bases de datos biométricos (edad, datos clínicos, información genética, datos ambientales, pruebas médicas) mediante aprendizaje automático va a incrementar la precisión del diagnóstico (ya hay avances al respecto en la detección del cáncer, patologías cardiovasculares y patologías neurológicas; Sánchez Caro & Abellán García Sánchez 2021).<sup>46</sup> Pero las predicciones basadas

---

46 El acceso a datos biométricos personales plantea la cuestión, que recogeremos al tratar la ética de la IA, de hacerlo con seguridad y respetando la privacidad (esto es, de forma que los datos estén anonimizados). También suscita la cuestión, que ya ha salido, del sesgo, pues, por ejemplo, los algoritmos de IA que diagnostican

en el descubrimiento de patrones estadísticos han de ser explicadas por la «ciencia médica» (no por la «ciencia de datos»), aportando explicaciones causales de los mecanismos biológicos involucrados en la enfermedad. Encontrar correlaciones en la jungla de datos no implica, empero, que medien relaciones de causalidad.

Un ejemplo estrella de correlación espuria en datos biomédicos nos lo proporciona el sistema de detección de neumonía compatible con la COVID-19 en que trabajaban Miriam Cobo Cano & Lara Lloret Iglesias (2023, 61-64). El sistema funcionaba muy bien, pero las investigadoras descubrieron, mediante la técnica de los mapas de calor, que a la hora de analizar una radiografía de tórax el algoritmo se fijaba más en la posición de los hombros que en la presencia de manchas en los pulmones. La razón era sencilla, aunque sorprendente: un paciente aquejado de neumonía tendía a hacerse la radiografía encogido de hombros. Desde luego, este patrón está correlacionado con tener neumonía, pero, de nuevo, correlación no implica causalidad (encogerse de hombros no es la causa de padecer neumonía sino, más bien, al revés, el efecto).

De otra parte, los sistemas con IA fallaron durante la pandemia de COVID-19 en la predicción de su curso, así como en el diagnóstico de la enfermedad, precisamente por la carencia de buenos conjuntos masivos de datos, como consecuencia de lo inesperado de la misma (Chakravorti 2022). Por otro lado, en febrero de 2013, Google Flu Trends predijo erróneamente más del doble de visitas al médico por síntomas relacionados con la gripe. Como señalan Lazer & *al.* (2014), este sonoro error mostró que el tamaño no importa, o por mejor decir, que no siempre se trata de usar *big data*. Pues unas veces –caso del COVID– podemos no tener una cantidad suficiente de datos y otras veces –caso de la gripe–, aunque tengamos muchos datos, estos pueden no ser representativos o relevantes (en ocasiones, añadir datos y más datos es como añadir epiciclos).

El aprendizaje automático sobre datos masivos no resuelve el problema de la inferencia causal en medicina: los algoritmos pueden ser muy buenos prediciendo resultados, pero los predictores no tienen

---

enfermedades de la piel cometen errores con personas de color negro o amarillo, como consecuencia de que las bases de datos contienen mayoritariamente imágenes de personas blancas.



por qué ser las causas. Además, las redes neuronales del aprendizaje profundo, que por su gran número de pesos ajustables son idóneas para aprehender distribuciones en espacios de alta dimensión ligadas a tareas complejas relacionadas con el cómputo de grandes conjuntos de datos clínicos etiquetados, pueden predecir correctamente qué le pasa al paciente, pero las variables intermedias que emplean para realizar el diagnóstico pueden ser ininteligibles, al no ser asimilables a las propiedades morfológicas en que se fija el médico y cuyo papel trata de comprender para intervenir con el tratamiento (Ríos Insua & Gómez Ullate 2019, 74). En otras palabras, las redes neuronales multicapa funcionan como «cajas negras», donde no siempre resulta fácil explicar por qué toman la decisión que toman. Y a ningún paciente le gustaría que le diagnosticaran sin que el médico pueda explicarle por qué tiene o deja de tener cierta enfermedad («el ordenador me ha dicho esto, pero no sé decirle por qué ha tomado esa decisión clínica sobre usted»).<sup>47</sup> La IA posibilita procesar más rápido y mayor cantidad de datos, pero puede conducir al reconocimiento de patrones irrelevantes o incomprensibles.

Además, como sugiere Aschwanden (2020), la IA, con su capacidad para arrojar correlaciones insospechadas, puede empeorar el sobrediagnóstico y el sobretratamiento (la hipermedicación) sobre falsos positivos, como consecuencia de que se empiece a tratar a un paciente sólo porque ha dado positivo en cierta correlación entre variables biomédicas.

En 2016, Geoffrey Hinton, padre del *deep learning*, predijo que los radiólogos serían sustituidos por IAs en menos de una década. En 2022, la Unión Europea ha dado luz verde al uso de un sistema con IA que examina las radiografías de tórax sin contar con un radiólogo, salvo cuando aquellas presentan anomalías o son dudosas, en cuyo caso son revisadas por uno (Cobo Cano & Lloret Iglesias 2023, 12). Transcurrido ya más de un lustro del vaticinio de Hinton, estamos más cerca de la simbiosis que de la sustitución. La IA no va a sustituir a

---

47 Una situación análoga se está dando en meteorología, donde los sistemas de predicción basados en IA, entrenados con datos meteorológicos de decenas de años, están obteniendo resultados similares a los modelos meteorológicos basados en ecuaciones físico-matemáticas que modelan el tiempo. Los modelos basados en IA son más rápidos, pero –por así decir– no saben nada del clima, simplemente establecen patrones.

los radiólogos, porque un radiólogo no sólo sabe clasificar imágenes –tarea que el aprendizaje profundo realiza estupendamente–, sino sobre todo razonar, en situaciones de ambigüedad, a tenor de la imagen médica, la historia del paciente y el conocimiento establecido, así como realizar diagnósticos y tomar decisiones en cuanto al tratamiento. A lo sumo, como sugieren Chan & Siegel (2019), la IA favorecerá una nueva generación de radiólogos con conocimientos en ciencia de datos (estadística).

Como apostilla Luis Martín Arias (2023), la medicina, a diferencia de la biología, no es una ciencia sino una praxis tecnológica, donde el saber hacer del médico está dirigido *in recto* a la individualidad corpórea del paciente, al paciente individual y concreto (con su ojo clínico). Por ello, noticias como la que se dio a conocer en mayo de 2023 respecto a que ChatGPT había pasado con éxito un examen de medicina de la Universidad de Harvard, olvidan la diferencia entre saber de biología y practicar la medicina. La IA es una herramienta auxiliar en el saber hacer con el paciente, a la hora de realizar diagnósticos (detección de melanomas, mamografías asistidas por IA para detectar nódulos cancerígenos), pronósticos (estimación del número de días que un enfermo va a pasar en un hospital de cara a optimizar la gestión asistencial) y tratamientos personalizados (el sistema AIDD usa IA para el descubrimiento de nuevos fármacos buscando correlaciones entre compuestos químicos y efectos sobre enfermedades, robots inteligentes usados en cirugía). Para Martín Arias (2023), serán aquellos médicos envueltos en la ideología tecnocrática que reduce la corporeidad del paciente a un mero dato estadístico los que, sin duda, podrán ser sustituidos por máquinas. Pero no lo serán aquellos médicos que mantengan el contacto con el cuerpo y la salud del paciente. La IA podrá automatizar algunas tareas médicas pero no absolutamente todas, de manera que más que ignorar o sustituir, se trata de caminar hacia una simbiosis entre médicos y máquinas.

Con estos mimbres, Larson (2022) mira a la abducción de Charles S. Peirce.<sup>48</sup> La abducción es un razonamiento apagógico (por decirlo

---

48 Este lógico y estadístico norteamericano (nacido en 1839 y fallecido en 1914), que pasó por España y cuya S. abrevió durante un tiempo el nombre de Santiago en español, resulta un filósofo en extremo interesante desde las coordenadas del

en términos aristotélicos), esto es, una clase de razonamiento –distinto de la deducción y de la inducción– por el que hacemos conjeturas ante un hecho sorprendente en función del contexto y la experiencia, tomando partido por una de ellas. Mediante la abducción no sólo podemos inferir la mejor explicación de un fenómeno, sino que también podemos realizar inferencias causales, seleccionando entre causas o factores enfrentados en la determinación de un efecto.

A pesar de que el proceso abductivo no se deja capturar formalmente ni mecanizar, podemos representarlo así:  $B, A \rightarrow B, \textit{ergo} A$ . En este caso, A constituye la explicación o la causa de B. Obsérvese que no estamos ante una deducción, porque se está cometiendo la falacia lógica de la afirmación del consecuente. Y tampoco estamos ante una inducción, porque B se supone que es un hecho nuevo y sorprendente, es decir, que no se repite con la suficiente frecuencia como para generalizar.

La abducción se emplea en ciencias como la física o la química para inferir explicaciones o causas de observaciones o efectos desconcertantes. Una ilustración la proporciona el descubrimiento del planeta Neptuno, como explicación y causa de las anomalías orbitales observadas en Urano. En matemáticas, la abducción aparece cuando el matemático investiga un problema cuyo enunciado no garantiza la existencia de respuesta y ha de ir afinando las diferentes alternativas, esto es, las hipótesis precisas para resolver total o parcialmente el problema. Y en medicina, la abducción hace acto de presencia cuando el médico diagnostica una enfermedad a partir de los síntomas observados en el paciente (síntomas que suelen variar de un paciente a otro y casi nunca son exactamente iguales).

Pues bien, mientras que los sistemas con IA lógica, como los sistemas expertos, eran esclavos de la extensión y la minuciosidad de las reglas codificadas en ellos, y los sistemas con IA basada en el aprendizaje automático, como las redes neuronales, son esclavos de la cantidad y la calidad del muestreo de entrenamiento en relación con la población

---

materialismo filosófico. No sólo por su sistematismo, la reivindicación que hacía de la escolástica o la importancia que concedió a las ternas en filosofía, sino especialmente porque en su sistema encontramos doctrinas coordinables –coordinar no es identificar– con el hiperrealismo o con el materialismo formalista propios de la teoría del cierre categorial. Un análisis de su concepción diagramática de las matemáticas puede leerse en Madrid Casado (2022a).

total, los seres humanos poseemos la capacidad de razonar –aparte de deductiva e inductivamente– abductivamente: con poca información, con tan solo unos cuantos ejemplos de muestra, logramos abstraer la regla general que opera detrás. Una red neuronal artificial necesita procesar un elevado número de imágenes de gatos, pero a un niño pequeño le basta con ver un gato una vez para saber reconocerlo. Esta versatilidad o flexibilidad humana, que nos diferencia de las máquinas lógicas y estadísticas, está condensada en el adagio latino *intelligenti pauca*, o sea, al inteligente pocos datos le son necesarios.

En verdad, en problemas donde se disponga de pocos datos, el *machine learning* no es una opción, dado que su fuerza (bruta) descansa en machacar muchísimos datos. Datos que, además, han de ser lo suficientemente variados, pues de lo contrario las generalizaciones resultantes serán deficientes, puesto que los algoritmos asumen un mundo estable, de modo que son buenos para interpolar entre datos conocidos, pero malos para extrapolar más allá de los datos de entrenamiento, cuando se presentan datos nuevos e inesperados.

Es cierto que algunos autores –como Herbert Simon– han sostenido que algunos programas informáticos, como BACON, son capaces de obtener leyes como la ley de Boyle, la ley de Ohm o la tercera de Kepler a partir simplemente de los datos, pareciendo que realizan abducciones. Pero conviene aclarar que lo que realmente hacen es inducciones, pues precisan tanto del suministro de numerosos datos como de ciertas reglas heurísticas: «determina si las mediciones son inversamente proporcionales y entonces busca constantes»; «si los valores de una variable son constantes, infiere que la variable es constante»; «si los valores de dos variables dan una línea recta, infiere que están relacionados linealmente»; &c. De hecho, estos programas no arrojan ninguna ley teórica novedosa, sino leyes precocinadas. Y la llamada programación abductiva es, en el fondo, programación lógica deductiva o programación bayesiana inductiva, donde todo está ya dado, considerándose una serie de hipótesis de partida y observándose cómo cambia la probabilidad de cada una de ellas en función de los datos (Larson 2022).

Es interesante que ChatGPT conteste, al preguntarle, que usa la deducción y la inducción, así como la abducción, aunque en menor medida y con capacidad reducida. De hecho, ChatGPT explica que

«la abducción implica partir de un hecho nuevo o inesperado que no puede ser explicado por las hipótesis o teorías existentes, y que requiere una explicación adicional». Pero tanto por la forma como por el contenido de la respuesta que da el chatbot cabe dudar de ese carácter novedoso que caracteriza a la abducción, ya que, al operar con textos ya escritos, se basa en teorías existentes, que parafrasea mediante una ingeniosa sustitución de sinónimos, &c. Por ejemplo, cuando se le pide que explique por qué hemos visto una nube roja cuadrada, adapta las teorías de la contaminación del aire, la posición solar o la manipulación digital al caso concreto. En resumen, lo que hace es deducción e inducción (extraer patrones a partir de los datos de entrenamiento). En sentido estricto, no hace abducción. «El aprendizaje automatizado no es más que una inducción automatizada» (Larson 2022, 160).

Ahora bien, para alcanzar la IA general, hace falta razonar abductivamente en situaciones nuevas nunca vistas, lo que precisa sentido común –algo, como vimos, en lo que ChatGPT, con sus habilidades sin comprensión, falla– y entender las conexiones causales –algo, como también vimos, que el mal llamado aprendizaje «profundo» no logra al confundir superficialmente correlación y causalidad– (Marcus & Davis 2014 y 2019). Y esto, como ampliaremos en el próximo capítulo, no se consigue mediante un diseño formal programado, en el que el programador proporciona a la máquina descripciones codificadas en un lenguaje de programación, sino operando materialmente en el mundo-entorno, pues va ligado a la corporalidad, razón por la cual Ramón López de Mántaras (2023) insiste en que no disponemos de ningún sistema que merezca el adjetivo «inteligente» sin comillas.

### **3.7 La IA realmente existente**

Hablamos de «IA realmente existente» a la manera que antaño se hablaba –recreando la consabida expresión de Mijaíl Suslov– del «comunismo realmente existente» (o del «cristianismo realmente existente») y hogaño se habla de la «democracia realmente existente». Pero no lo hacemos al objeto de contraponerla a una «IA ideal»

(fuerte y general), a la que la «IA actual» (débil y restringida) tendería a aproximarse, al modo en que hace décadas muchos pensaban que el comunismo realmente existente era una aproximación perfectible del comunismo ideal o, en nuestros días, muchos piensan que la democracia realmente existente dista del ideal de democracia y creen que los déficits de las democracias realmente existentes se solucionarán con más democracia.

Nuestra perspectiva es, precisamente, la contraria. Ese modelo ideal, límite y utópico, de IA fuerte y general, como de comunismo final, cristianismo eterno o democracia perfecta, no existe, asemejándose al *perpetuum mobile* de la termodinámica. Y no es que los motores reales sean aproximaciones más o menos lejanas al móvil perpetuo, como si este fuera un punto atractor alcanzable por la tecnología, sino que el concepto de móvil perpetuo cristaliza *ex post facto*, cuando se ordena la serie de motores reales constatando el límite de los rendimientos. El motor P será más eficiente que el motor Q, pero ni P ni Q pueden franquear cierto límite del rendimiento. Análogamente, la IA realmente existente es la IA débil y restringida que tenemos y que tan buenos servicios nos presta; porque no existe otra, porque la IA fuerte y general pide una y otra vez el principio. Pero, entonces, ¿por qué considerar la IA realmente existente como menos potente cuando, de facto, es la única real?

Considerar los avances en la IA del presente en marcha como pasos encaminados hacia la IA fuerte y general (como ha hecho Microsoft con GPT-4) es equiparable a sostener, por decirlo con Hubert Dreyfus (1965, 17), que el primer hombre que trepó a un árbol estaba progresando en la conquista de la Luna. Estamos tan lejos (o tan cerca) de la IA fuerte o general como lo estábamos hace setenta años, en los albores del campo.

## Capítulo 4

### *La IA no es inteligencia ni es artificial*

Sin restar fuerza a los argumentos contra la viabilidad de una IA fuerte y general expuestos en el anterior capítulo, nos parece que todos ellos se mueven a una escala epistemológica (sujeto/objeto), siendo preciso desbordarla y asumir una escala gnoseológica (materia/forma). Así, dejaremos de hablar de un sujeto o un objeto que comprende, razona o practica la deducción, la inducción o la abducción, para comenzar a hacerlo de conjuntos estructurados de sujetos y objetos, es decir, de instituciones.

El quid de la cuestión es que la IA no es, en puridad, inteligencia ni artificial, si por inteligencia se entiende algo parecido a la humana y si por artificial se entiende algo contrapuesto a natural (como suponen que es la inteligencia humana). Así lo ha aventado Kate Crawford (2021, 8): «La IA no es ni artificial ni inteligencia». O también Luc Julia (2019): «La IA no existe». Pero entre bastantes profesionales del sector persisten dos mitos firmemente establecidos: el mito de la inteligencia y el mito de lo artificial.

#### **4.1 El mito de la inteligencia**

Tenemos, por un lado, el mito de la «inteligencia separada». Inteligencia es, según el diccionario de la Real Academia Española, la

«capacidad de entender o comprender», «de resolver problemas». En sintonía, el etólogo P. J. B. Slater (2000, 224) define: «Inteligencia: Capacidad que permite a un individuo aprender tareas, razonar y resolver problemas».

Y, dentro del campo de la IA, Wang (2019, 17) define específicamente la inteligencia como la capacidad de un sistema procesador de información de adaptarse a su entorno mientras opera con conocimiento y recursos insuficientes (una definición que ajusta como un guante a la moderna IA guiada por los datos, la del aprendizaje automático, pero deja fuera a la vieja y buena IA simbólica, la basada en la lógica).

La dificultad estriba en que la noción de inteligencia aparece en múltiples dominios categoriales. Aparece en psicología (los test y los tipos de inteligencia humana), etología (la inteligencia animal, de los simios, los elefantes, los delfines o las ratas), sociología (la *intelligentsia*) e, incluso, política (los servicios de inteligencia). Esta tesitura invita a pensar que estamos ante una idea filosófica más que ante un concepto científico, categorial (para la distinción entre conceptos e ideas, véase 2.1).

En efecto, para el Grupo de Expertos de Alto Nivel sobre IA de la Unión Europea, en su documento *Una definición de la inteligencia artificial: Principales capacidades y disciplinas científicas*:

El término IA contiene una referencia explícita al concepto de inteligencia. Ahora bien, dado que la inteligencia (sea en las máquinas o en los seres humanos) es un concepto vago pese a haber sido estudiado en profundidad por psicólogos, biólogos y neurocientíficos, los investigadores que trabajan en el terreno de la IA utilizan principalmente el término «racionalidad». La racionalidad hace referencia a la capacidad de elegir la mejor acción posible para alcanzar un objetivo determinado, dados determinados criterios que es necesario optimizar y teniendo en cuenta los recursos disponibles. Por supuesto, la racionalidad no es el único ingrediente del concepto de inteligencia, pero sí uno muy importante (Grupo de Expertos de Alto Nivel sobre IA de la Unión Europea [GEANIA] 2019b, 1).

Pero con la noción de racionalidad ocurre lo mismo que con la noción de inteligencia. No existe una definición universalmente aceptada de



razón, como no la hay de inteligencia, porque no hay un único criterio sino dispares. No es un concepto científico sino una idea filosófica, por cuanto de nuevo aparece en múltiples categorías: matemáticas (la razón entre dos magnitudes, las razones trigonométricas, los números racionales), lógica (razonamiento deductivo e inductivo), psicología (la racionalidad humana), etología (la racionalidad animal), economía (la teoría de la elección racional), política (la razón de Estado).

Lo que aquí nos interesa destacar es que los profesionales del sector de la IA tienden a entender la inteligencia y la racionalidad en términos puramente formales, al margen de los materiales sobre los que se realiza. Así, suelen maravillarse de las «obras de arte» que crea DALL-E, obviando que lo que produce no son pinceladas sino píxeles (como decía Lucian Freud: «El ingrediente sin el cual la pintura no puede existir: PINTURA»), o sea, pigmentos, costras de empaste, &c.). Al hacerlo así, tributan sin quererlo al dualismo cartesiano más metafísico. La dicotomía mente/cuerpo (*res cogitans/res extensa*) reaparece ahora como *software/hardware*, reduciéndose la inteligencia a operaciones mentales (formales), no corporales (materiales), pues el soporte resulta secundario, dando igual que sean neuronas orgánicas o chips de silicio. De hecho, para el ingeniero de Google Blaise Agüera y Arcas (2021), «los enormes modelos de lenguaje ilustran por primera vez la forma en que la comprensión del lenguaje y la inteligencia pueden disociarse de todas las características corporales y emocionales que compartimos entre nosotros y con muchos otros animales».

Este reduccionismo se acerca a la idea formalista de razón de la tradición escolástica, que entendía la razón como el hábito de sacar conclusiones, esto es, como la capacidad de hacer silogismos, vinculada a la deducción lógica (una capacidad que no era exclusiva de los hombres, pues animales racionormorfos como el perro de San Basilio también resolvían silogismos disyuntivos). Hoy, sus partidarios extienden la definición escolástica haciendo referencia a la capacidad de aplicar los algoritmos inductivos del aprendizaje automático, vinculados al cálculo de probabilidades y la estadística. Para Kaplan (2017, 7), por ejemplo, la esencia de la inteligencia es «la capacidad para hacer generalizaciones adecuadas de un modo oportuno, basándose en datos limitados».

Pero, como apunta Crawford (2021, 7), la visión de la mente como una computadora y la visión de la computadora como una mente

desencarnan la inteligencia, convirtiéndola en un mero formalismo (lógico o estadístico). Estas visiones rozan el cerebrocentrismo, que encapsula la inteligencia en operaciones formales desplegadas por redes neuronales, sean de carbono o de silicio. Desde nuestras coordenadas, es preciso reencarnar esta concepción de la inteligencia y de la razón, sustituyendo el entendimiento o la mente por la corporeidad operatoria del sujeto.<sup>49</sup> Se trata de anteponer lo práctico a lo teórico. La razón práctica es la razón originaria, y la razón especulativa es su resultante. La referencia filosófica ineludible no es el cogito cartesiano o el entendimiento kantiano sino el sujeto corpóreo operatorio (Bueno 1992, 439).

El *logos*, tal como lo entendemos desde el materialismo filosófico, no es una facultad subjetiva (espiritual o cerebral), porque es inseparable del cuerpo orgánico, va ligado a las operaciones quirúrgicas, a juntar y separar con las manos, aun cuando cambien los instrumentos. El *logos* es llevado a efecto por operaciones manuales antes que meramente verbales, tejiendo o ensamblando unas cosas con otras, componiendo materiales (Bueno 2008 y 2011). (Hay que observar que la propia tecnología computacional necesita, antes que nada, manipular estados de dispositivos electrónicos.) El materialismo filosófico está más cerca en este punto de los etólogos, que hablan del chimpancé Sultán como un animal racionomorfo capaz de componer herramientas para alcanzar alimentos –aun cuando no se trate estrictamente de técnicas al no estar normadas–, así como de los antropólogos, que hablan del *homo sapiens* como indisociable del *homo habilis* y del *homo faber* (Bueno 1992, 111).

Frente a la hipertextualización dominante, hay que incidir en que no sólo existe el conocimiento textual, recogido en documentos, como los digitalizados de que se nutren ChatGPT u otros programas (por ejemplo, los programas de reconocimiento de imágenes precisan de un aprendizaje supervisado en el que durante el entrenamiento se etiquetan –se textualizan– las imágenes, escribiendo «esto es gato» o «esto no es gato», como dice Cervantes que hacía el pintor

---

49 El atinado diagnóstico de Kate Crawford se explica porque es deudora –como reconoce– de los trabajos de Ian Hacking, Peter Galison o Lorraine Daston (Crawford 2021, 12), cuya confluencia con Gustavo Bueno hemos analizado en Madrid Casado (2018a, cap. 12).

Orbaneja debajo de sus cuadros). No todo el conocimiento está digitalizado ni es digitalizable. Existe una gran masa de conocimiento ligado a la operatoriedad corpórea, esto es, a la manipulación quirúrgica de objetos extrasomáticos. Los seres humanos no sólo aprendemos de leer textos o contemplar imágenes (como si sólo fuéramos un cerebro que recibe inputs y emite outputs a la manera de un ordenador), sino fundamentalmente de hacer cosas con cosas. Las correlaciones son digitalizables; pero las conexiones causales no. Los mecanismos relacionados con la causalidad y, en general, el cúmulo de conocimientos fruto de nuestras operaciones con el entorno desde la infancia (lo que se suele abreviar como «sentido común») dependen de un sujeto gnoseológico concebido como sujeto operatorio corpóreo. Los sistemas con IA son, como decíamos en el capítulo anterior, sistemas basados en conexiones paratéticas; pero el ser humano despliega relaciones apotéticas, a distancia, que tienen que ver con las relaciones entre el cuerpo y el entorno, con la finalidad propositiva, consistente en transformar rotacionalmente el allí-entonces en aquí-ahora (*hic et nunc*).

Abundando: la relación del hablar con el manipular ya fue señalada por Platón en el *Crátilo*. Hablar es fundamentalmente juntar y separar sonidos, moviendo los músculos estriados de nuestro aparato fonador, de modo semejante a como movemos las manos para componer o descomponer cosas corpóreas. Y hablar también es gesticular, que es lo que, entre otras cosas, dota de referenciales apotéticos al lenguaje (Bueno 1985a). (Aún más, si estamos chateando por Internet, son nuestros dedos los que hablan.) Frente a Chomsky, algunos como Kaplan (2017, 67) parecen volver a Skinner al señalar que el lenguaje surgió como una extensión natural de los gestos: los hechos con la lengua y la boca en lugar de con las manos y los brazos.

En suma, «el hombre piensa porque tiene manos», como decía Anaxágoras (o, más bien, como dice Aristóteles que dice Anaxágoras en *De partibus animalium*, IV, 687a 8-9). Un anclaje manual de la inteligencia, que la empapa de experiencia, de lo que los dedos tocan, acercan o separan, y no la hace previa ni independiente del mundo entorno que manipula (literalmente) y, de paso, articula. La racionalidad tiene tanto que ver con las manos como con los ojos, la laringe o los oídos (Bueno 1995a, 39-40).

Sentado esto, el materialismo filosófico reformula la inteligencia como la capacidad para neutralizar contradicciones (los famosos «problemas»), y la razón estaría no tanto en la facultad para hacerlo como en el propio proceso y producto. La racionalidad no es algo sobreañadido al dominio humano, y que los humanos pusiéramos como aderezo gracias a nuestras mentes racionales, sino la misma composición operatoria –técnica, científica y tecnológica– de los materiales del mundo entorno por parte de los sujetos corpóreos (Bueno 2005). Al operar, los seres humanos nos topamos con contradicciones, porque lo que mayormente se resisten son los objetos, no las palabras (no puedes meter un objeto cuadrado en uno triangular, pero sí puedes pronunciar «círculo cuadrado»). Bueno (2005, 25) describe el procedimiento operatorio de neutralización de una contradicción como un arco de racionalidad que comprende tres fases: proposición, contraposición y recomposición. Pero las racionalidades siempre son múltiples (razón matemática, física, biológica, económica, política, &c.), de modo que la irracionalidad no es la ausencia de toda racionalidad (como si eso fuera posible) sino la discordancia entre dos racionalidades inconmensurables e irreductibles entre sí.

Resumiendo: mitologizar la inteligencia, convirtiéndola en una suerte de sustancia inmaterial, descorporeizada, es engañoso. Una suerte de refluencia teológica. Por decirlo a la contra: no puede haber inteligencia general sin cuerpo, no hay inteligencias separadas (sean angelicales o virtuales, estén en el Cielo o en el Metaverso).

Entonces, cuando se toma el test de Turing como criterio (formal, no material) de inteligencia y racionalidad, de manera que si un programa de ordenador lo pasa, se deduce que es equiparable a un ser humano, se está cometiendo una reducción:

El principal efecto de esta reducción es, a nuestro juicio, que la idea de persona queda segregada de la individualidad con su morfología humana característica (por ejemplo, de las manos o del aparato fonador). [...] No se trata de postular que, en un momento dado, el ordenador de referencia emane un «alma», o una «conciencia», que en realidad, sería un simple epifenómeno que no intervendría para nada en el curso de las operaciones de la persona ordenador. Lo que necesitamos postular es que el ordenador, en el supuesto momento de cumplir el criterio de

Turing, desarrolle un cuerpo humano, puesto que sólo a partir de ese cuerpo (de sus manos, de su aparato fonador, &c.) podrá desarrollar su *logos*, su razón, su conciencia. Pero como esto es imposible, habrá que considerar la definición [...] como el ejercicio de una mera analogía entre un ordenador y unas personas dotadas de lenguaje; una analogía acompañada de una hipótesis que pide el principio que trata de demostrar, a saber, que un ordenador sin figura humana puede satisfacer el criterio de Turing (Bueno 1996b, 140).

Y Bueno (1996b, 163) añadía con palabras que casi tres décadas después siguen de permanente actualidad:

No basta definir al hombre como animal racional, salvo que se dé por descontado que este animal racional ha de tener la forma humana: por mucho que un loro pronuncie discursos racionales (decía Locke) nunca llegará a ser considerado como hombre. Más aún, por mucho que un ordenador resuelva problemas algebraicos, traduzca lenguas extranjeras o mantenga «conversaciones», no podrá ser considerado como una persona humana, y no ya porque carezca de «conciencia», «sentimiento» o «alma», sino porque carece de cuerpo humano.

Sustentándose en la tradición fenomenológica, Hubert Dreyfus (1992) también ha criticado la razón artificial no corpórea y ha insistido en que la corporalidad desempeña un papel crucial para la inteligencia. Y López de Mántaras (2018, 166-167) advierte que una de las críticas más fuertes a la visión incorpórea de la IA pasa por reconocer que la inteligencia necesita de cuerpo, ya que depende inextricablemente de las capacidades sensoriales y motoras, de la interacción con el entorno (cognición situada). En palabras de López de Mántaras & Meseguer González (2017, 14-15):

Una de las críticas más fuertes a estos modelos no corpóreos [de la IA] es que un agente inteligente necesita un cuerpo para poder tener experiencias directas con su entorno [...] Sin un cuerpo, estas representaciones abstractas no tienen contenido semántico para la máquina [...] En otras palabras, *el cuerpo da forma a la inteligencia* (*the body shapes the way we think*) y por lo tanto sin cuerpo no puede haber inteligencia de tipo general.

A su entender, los avances más prometedores en IA pasarían por la robótica que persigue encarnar el *software* avanzado en agentes equipados con sensores que les permitan interactuar con el mundo entorno, integrando la percepción, la representación, el razonamiento, la acción y el aprendizaje (López de Mántaras y Meseguer González 2017, 149-150).

Pero hay más. Como observa Alfredo Marcos (2020), las tareas y los problemas, las contradicciones, lo son siempre para nosotros, pues somos los seres humanos los que queremos realizar tal cómputo o alcanzar tal meta. En ningún caso lo son para las máquinas, que carecen de intencionalidad, de prolepsis y conducta apotética. Se habla de *machine learning*, de aprendizaje maquina, pero decir que las máquinas aprenden es, en el fondo, una antropomorfización, porque somos nosotros los que aprendemos con ayuda de las máquinas, con su asistencia. Somos nosotros los que tenemos el problema, preparamos los datos, seleccionamos y entrenamos el algoritmo, lo ejecutamos e interpretamos los resultados. Si una máquina realiza un clúster con datos médicos, es porque hay un grupo de médicos interesados en ello, no porque la máquina quiera, ya que el sentido de la construcción está en función de ese grupo de médicos y de los pacientes implicados. Marcos (2023) emplea este quiasmo: la parte inteligente del sistema no es artificial y la parte artificial del sistema no es inteligente. La IA no es inteligente *per se*. Su inteligencia es la que sus inventores han depositado objetivamente en ella. (Por consiguiente, como veremos más adelante, la responsabilidad ética sigue siendo de los agentes humanos –programadores, ingenieros y usuarios–, no de los propios sistemas de IA.) Y es que para Bueno (1985b, 107):

No es un hombre individual el que ha fabricado la computadora, sino un hombre social e históricamente determinado; por así decirlo, la estructura de la computadora reflejaría no ya un cerebro, sino un conjunto de cerebros, y un conjunto de cerebros no es un cerebro, como un conjunto de poliedros regulares no es un poliedro regular.

La IA ha de estar inserta en una red que incluya humanos para funcionar y tener sentido. A la manera que Bueno (1996a) decía que «una televisión que no es vista por nadie no funciona como tal, sólo

si es vista por alguien hay mensaje o información», podemos afirmar que un sistema con IA que no es usado por nadie o no afecta a nadie no funciona como tal, sólo si es usado por o afecta a alguien hay inteligencia. Esto es clave. Porque esa red grupo-máquina-grupo nos pone sobre la pista de las instituciones.

Tanto frente a los que confinan la inteligencia o la racionalidad a la mente o al cerebro humanos, como frente a los que ponen la inteligencia o la racionalidad en los ordenadores al margen y con independencia de los hombres –lo que Marcos (2023) denomina «efecto *Toy Story*», por los juguetes que cobran vida cuando el niño no juega con ellos–, el materialismo filosófico sustenta, rectificando dialécticamente ambos extremos, que la racionalidad subjetual y la racionalidad objetual se dan siempre conjugadas (Bueno 2005, 16). El ingenio subjetivo se demuestra por los ingenios objetivos contruidos, sin perjuicio de que estos precisen a su vez del ingenio subjetivo para ser armados. Según Bueno (2005, 16):

La barrera que impide, o dificulta, la ampliación de la idea de institución racional al terreno de las estructuras corpóreas extrasomáticas, tales como ordenadores, edificios, &c., está formada, con toda seguridad, por los prejuicios espiritualistas reforzados en la versión del dualismo cartesiano; prejuicios que obligan a reducir la racionalidad al campo de un metafísico sujeto espiritual.

Pero el rechazo del formalismo de la racionalidad tampoco ha de hacernos recaer –como explica Bueno (2008, I.A)– en el materialismo sustancialista de la racionalidad, que, a diferencia del materialismo actualista, entiende la racionalidad como un atributo intrínseco de ciertas entidades materiales (sean los ordenadores o, como antaño, la Naturaleza). La razón, la inteligencia, tal y como las entiende el materialismo filosófico, no son previas a las operaciones, sino que se dan *in medias res*, a través de las operaciones con objetos extrasomáticos (la memoria, por ejemplo, no se reduce a una facultad cerebral, puesto que depende de la escritura). Se forman históricamente, en el proceso de constitución de instituciones (en concreto, tecnológicas). El hombre es el animal institucional por excelencia (Bueno 2005, 8) o, si se prefiere, transformando la

definición lisológica de hombre en morfológica, el hombre es el animal racional en tanto que es racional por las instituciones (lo que da la diferencia específica entre la racionalidad humana y la conducta racionomorfa de los animales; Bueno 2005, 24). Y esto nos pone a las puertas del segundo mito.

## 4.2 El mito de lo artificial

Suele considerarse la inteligencia humana como natural (como si residiera en los genes), en contraposición a la inteligencia de las máquinas, que sería artificial. Se constata, por tanto, la persistencia de los mitos de la Naturaleza y de la Cultura (Bueno 2016), funcionando a toda máquina en la conceptualización de la IA. Pero resulta que la inteligencia humana es ella misma artificial, pues no puede divorciarse de instituciones como la educación, los libros o las mesas. Para Crawford (2021, 5), la inteligencia no es algo natural, independiente del contexto social, cultural, histórico y político.

No es sólo que la inteligencia humana sea abductiva, situacional y operacional, como va visto, sino que está extendida («mente extendida»), externalizada en ingenios objetivos –las instituciones– que trascienden positivamente el ingenio subjetivo de cada individuo.<sup>50</sup> El materialismo filosófico

requiere la ampliación de la extensión de la idea de institución racional al campo radial de la cultura extrasomática, desbordando el campo conductual y sociológico (circular y angular) al que tradicionalmente viene circunscrita [...] La racionalidad, en su especificación de racionalidad humana, tampoco se predica única y directamente de las instituciones conductuales (por ejemplo, de las ceremonias), o de las instituciones sociológicas (por ejemplo, de las empresas financieras globalizadas); también se predica directamente de los ordenadores, de los templos, de los libros, de las ametralladoras o de los misiles (Bueno 2005, 16).

---

50 Por esto, la objeción de que la IA también es social (por las redes de ordenadores conectados) es falaz, ya que ahí la inteligencia no es realmente extrasomática, no está fuera del conjunto de CPUs concebido como un conjunto de cerebros flotantes.



Por consiguiente,

una ametralladora, o un misil, es una institución tan racional como pueda serlo un silogismo [...] por supuesto, la racionalidad de las instituciones extrasomáticas se manifestará en la concatenación de sus partes involucradas con la racionalidad conductual de los sujetos que las utilizan (Bueno 2005, 25-26).

Esta doctrina mantenida por el materialismo filosófico no es extraña a filósofos como Charles S. Peirce, Bruno Latour o Javier de Lorenzo. Peirce sugirió que la inteligencia humana es artificial porque depende del uso de artefactos. En matemáticas, del uso de diagramas y grafos. En química, Lavoisier hizo de sus alambiques y aparatos instrumentos de pensamiento, alcanzando una nueva forma de razonamiento como algo que manipula cosas reales en lugar de palabras y fantasías (CP 5.363). Y, en su opinión, «es más cierto que los pensamientos de un escritor vivo están antes en cualquier copia impresa de su libro que en su cerebro» (CP 7.364). Por su parte, para Latour (1996, 300-1):

Con la introducción de tantas tecnologías intelectuales, desde la escritura hasta los laboratorios, desde las reglas hasta los guijarros, desde las calculadoras de bolsillo hasta los entornos materiales, se ha desdibujado la propia distinción entre las inteligencias naturales, situadas y tácitas, y las artificiales, transferibles y desencarnadas.

Es aquello que Javier de Lorenzo (2020, cap. 1) explica al decir que los seres humanos habitamos un mundo de artefactos materiales (como casas, ordenadores, sillas, mesas...), conceptuales (como los elementos geométricos, aritméticos, estadísticos o computacionales que estructuran el hábitat en que vivimos, nos movemos o pagamos con dinero) y simbólicos (como las ideologías).

En definitiva, la inteligencia no está dentro de nuestras cabezas, pero tampoco está ahí afuera con independencia de nosotros. La inteligencia y la racionalidad están en nuestras manos, en nuestros cuerpos, que las diseminan por el mundo entorno conforme lo transforman constituyendo instituciones técnicas, tecnológicas, &c.

¡Tanta IA tiene un libro de matemáticas como un ordenador! Tanta «inteligencia artificial» hay en un teorema matemático o una sonata como en ChatGPT o Dall-E.

Parte III

ÉTICA DE LA  
INTELIGENCIA ARTIFICIAL



## Capítulo 5

### *Ética, moral y política de la IA*

En capítulos anteriores hemos mostrado que la IA es una idea-fuerza muy oscura y confusa, cuyas inconsistencias gnoseológicas y ontológicas hemos tratado de sacar a la luz. El mito de la IA consiste, como va visto, en que lo que se designa por tal nombre no es propiamente inteligencia ni específicamente artificial. En esta tercera y última parte, abordamos la llamada «ética de la inteligencia artificial». Pero, ¿de qué hablamos cuando hablamos de ética de la IA y de los llamados «problemas éticos» que plantean ChatGPT, MidJourney, los asistentes personales, las aplicaciones en medicina o el uso que estos sistemas hacen de nuestros datos?

Desde las coordenadas del materialismo filosófico, mejor que de ética de la IA (y de problemas éticos) debe hablarse –como demostraremos– de ética, moral o política de la IA. Estos tres planos (ética/moral/política) intersecan en lo que suele denominarse como «ética de la inteligencia artificial» y aparecen frecuentemente confundidos (como comprobaremos al analizar, entre otras, la normativa para una «IA ética» promovida por la Unión Europea), pese a que resulta imprescindible diferenciarlos en el momento de tomar decisiones relacionadas con los riesgos que plantea la IA: la invasión de la privacidad, el sesgo en los resultados o la responsabilidad por los errores. Concluiremos llamando la atención sobre la importancia de la geopolítica en el control de la IA.

## 5.1 Tecnoética y bioética

El rótulo «tecnoética» fue acuñado por Mario Bunge (1977), aunque los principales contribuyentes al despegue de la llamada «ética aplicada a la tecnología» han sido el filósofo alemán Hans Jonas y el filósofo norteamericano Carl Mitcham. Como consecuencia de la inflexión que supone la expansión acelerada de múltiples tecnologías *bio*, *nano* e *info*, la tecnoética constituye actualmente un campo de investigación bastante vivo (Luppicini & Adell 2009). Para Echeverría (2010, 144), las tres principales ramas de la tecnoética son la bioética (la primera en surgir desde un punto de vista cronológico), la nanoética y la infoética. La ética de la tecnología comprende la ética de la IA, pero también la «ciberética» (la ética de Internet y de otros entornos digitales), así como la ética de los ingenieros y del resto de profesionales involucrados en el desarrollo o el uso de tecnologías, el estudio de su implementación social y la valoración de impactos y responsabilidades.

Es un lugar común afirmar que, mientras disponemos de una bioética madura (¡como si hubiera una sola!), no tenemos todavía una tecnoética o una ética de la tecnología –y, en concreto, de la IA– plenamente desarrollada (Diéguez 2017, 62). Por ello, la «comunidad ética» y diversas organizaciones deberían volcarse en elaborar una tecnoética y, en particular, una IA-ética, pues (supuestamente) «la fórmula de los Comités de Bioética puede ampliarse sin dificultades a los Comités de Tecnoética, en los que pueden integrarse personas de muy diferentes formaciones académicas y profesionales, siempre que tengan que afrontar y resolver problemas morales suscitados por el uso de las diversas tecnociencias en sus respectivas actividades profesionales» (Echeverría 2010, 145).

El quid de la cuestión es que no existe una única tecnoética, al igual que no existe una única bioética; porque ni la tecnoética ni la bioética son ciencias. Pese al armazón institucional (cátedras, sociedades, revistas...) que busca homologar la bioética con otras disciplinas científicas, los comités bioéticos no son comités científicos, por más que estén formados por médicos, biólogos o ingenieros. No existe la Bioética, como no existe la Tecnoética (así, con mayúscula y en singular). Hay múltiples bioéticas y tecnoéticas (con minúscula y en plural). Porque no hay una unidad doctrinal: los consensos de los comités éticos encargados de evaluar propuestas relacionadas con la medicina, la biología o las

llamadas «nuevas tecnologías» pueden ser fruto de principios distintos, que confluirán en unos casos y en otros casos entrarán en conflicto, siendo incompatibles unas posturas con otras (como dijo Jacques Maritain en la Comisión de Derechos Humanos de 1947: «podemos estar de acuerdo en muchas cosas, con tal de que no nos pregunten por las razones»).

Ante la tecnoética, como ante su hermana mayor, la bioética, estamos delante de una disciplina pragmática, definida por un conjunto de problemas relacionados con las tecnologías, pero cuyo campo acusa una unidad precaria, polémica, dado que está atravesado de conceptos científicos pero también de ideas filosóficas. En consecuencia, siempre que se hable de tecnoética o de ética de la IA hay que dar el apellido. Contra el eclecticismo, hay que especificar a qué tecnoética o a qué ética de la IA nos estamos refiriendo.

La tecnoética apunta a la gestión de la tecnología por el hombre. Ahora bien, a la hora de determinar el objeto de esa gestión hay que distinguir dos grandes corrientes, de la misma manera que ocurre con la bioética (Bueno 2001). Por un lado, tenemos la corriente que toma como objeto la vida humana. Las *tecnoéticas antrópicas* se centran específicamente en la vida humana, porque la gestión de la tecnología sólo puede hacerse desde una parte, desde la parte antrópica (sin perjuicio de la no maleficencia de animales ni vegetales, en tanto su subsistencia repercute en la de la vida humana). Y, por otro lado, la que toma como objeto la vida en general, la biosfera. Las *tecnoéticas anantrópicas*, por influencia del animalismo, el antiespecismo y el ecologismo, toman como referencia todas las formas de vida. Así, en su libro *Ética de la inteligencia artificial*, Coeckelbergh (2021, cap. 12) llega a cuestionarse si la ética de la IA ha de ser antropocéntrica, dada la influencia de la IA en otros seres vivos y en el medio ambiente. Y, en *La filosofía política de la inteligencia artificial*, Coeckelbergh (2023, cap. 6) considera no sólo una ética sino una política anantrópica, que se focaliza en formas de vida no humanas. Otros autores incluso hablan sin precaución alguna de «ética de los robots» (de los cuales sólo una fracción minúscula posee forma humana, careciendo de ella la mayoría, como los coches o los aviones autónomos), fijándose en las máquinas como virtuales sujetos éticos.<sup>51</sup>

---

51 Esta clasificación interna y material (yendo a los contenidos) de las éticas en antrópicas y anantrópicas es mucho más potente que la habitual clasificación externa entre éticas religiosas y laicas (así, el manual de bioética de Cambridge, editado por

Sentado que no existe una única tecnoética sino varias, con premisas diferentes e incompatibles entre sí, es necesario tomar partido. Lo que no implica sectarismo, sino el reconocimiento de que no cabe enjuiciar determinadas cuestiones neutralmente, sin comprometerse dialécticamente por una alternativa opuesta a otras. En nuestro caso, tomamos partido por una tecnoética materialista, que se alinea en la dirección antrópica, considerando como metafísicas y místicas las opciones anantrópicas, por cuanto la gestión de la tecnología y, en particular, de la IA sólo puede llevarse a cabo desde una parte de la biosfera, desde el hombre. Por ejemplo, las tres leyes de la robótica propuestas por Isaac Asimov en un cuento de 1942 son importantes en tanto en cuanto toman como objeto a la vida humana.<sup>52</sup> Cuando se habla de ética de los robots, se está hablando de ética entre hombres a través del manejo de robots. Es así que la Unión Europea apuesta en su normativa por un «enfoque centrado en el ser humano» de la IA.

Pese a su apariencia, la exposición que vamos a hacer de la tecnoética materialista responde más a una estructura estromática que a una estructura axiomática. No es que partamos de los principios que vamos a explicar como si fueran autoevidentes, antes de progresar hacia los casos de estudio relacionados con la IA, sino que esos principios se han formado a partir del regreso desde casos concretos de estudio. Con otras palabras, no entendemos la distinción entre principios y reglas en un sentido proposicionalista, donde los principios son *a priori* y generales, y las reglas son *a posteriori* y particulares. Desde la teoría del cierre categorial (Bueno 1992), hay que entender la distinción entre principios y reglas al revés (Bueno 2001, 61 y sigs.). El punto de partida siempre son las reglas, dadas *in medias res*. Una regla es la representación de

---

Peter Singer, se articula en bioética católica, protestante, islámica, judía, aborigen, de los testigos de Jehová, &c.; Singer 2009, sección IX), y que las clasificaciones meramente formales entre éticas principialistas y consecuencialistas, o deductivistas e inductivistas (esta última distinción fundamentalismo–decisionismo es la que suele explicar el experto en bioética Diego Gracia sin reparar en que es genérica y no toma en cuenta los contenidos; Gracia 2007, Parte II).

52 Estas son las tres leyes de Asimov: 1) Un robot no debe dañar a un humano o, por inacción, permitir que un humano sufra daño alguno. 2) Un robot debe obedecer las órdenes dadas por los seres humanos, excepto cuando tales órdenes entren en conflicto con la primera ley. 3) Un robot debe proteger su propia existencia siempre y cuando dicha protección no entre en conflicto con la primera o la segunda de estas leyes.



una norma, es decir, la representación de la concatenación de una serie de operaciones en función de un fin como consecuencia de haberse convertido en una rutina victoriosa (Bueno 2001, 65-66). Pero sobre las reglas pueden estar actuando principios variados. Principios que, pese a su nombre, aparecen al final, porque son abstractos, y que, más que fundamentar, coordinan conjuntos de reglas. Los principios no admiten excepciones, mientras que las reglas sí («la excepción que confirma la regla», suele decirse). Los principios van ligados a términos y relaciones, mientras que las reglas lo están a las operaciones.

Ahora bien, toda tecnoética y toda ética de la IA ha de distinguir entre tres clases de reglas o normas prácticas: las normas éticas, las normas morales y las normas políticas (jurídicas). Por esta razón, la tecnoética es también tecnomoral y tecnopolítica.

Frecuentemente, en la estela de José Luis López Aranguren, Javier Muguerza y sus discípulas, se entiende la ética como el estudio o tratado de la moral, como si sólo los profesores de filosofía especialistas en la materia supieran de ética, lo que conduce a un elitismo absurdo que hace de estos una suerte de curas laicos (la «comunidad ética»), olvidando que para conducirse de modo ético no hace falta ser profesora o catedrática de ética, porque cualquier ciudadano que no sea un débil mental se plantea cuestiones éticas, morales y políticas (todos somos filósofos, aunque hay muchas filosofías). Además, desde este enfoque, la ética termina conformando una suerte de teoría proposicional susceptible de ser aplicada a diferentes campos (bioética, tecnoética, nanoética, infoética, ética aplicada a la IA, &c.). Pero el sintagma escolar «ética aplicada» está malformado, pues la ética no es un sistema exento susceptible de ser aplicado, puesto que la ética consiste en su misma aplicación o ejercicio, que, por descontado, implica ideas (Bueno 2001, 67).

En otras ocasiones, se entiende por ética el conjunto de reglas o normas prácticas dictadas por mi propia conciencia (normas autónomas), y por moral el conjunto de normas que no vienen de dentro sino de fuera, de la coacción de la iglesia o la presión de la sociedad (normas heterónomas). Actuar éticamente sería, según esto, actuar en conciencia, sin influencias externas. Pero esta concepción de raigambre kantiana desemboca en el relativismo ético más radical, por cuanto se apoya en la conciencia subjetiva, de modo que habría

tantas éticas como conciencias se erijan en fuente de normas éticas. Además, esta doctrina concibe la conciencia individual como algo prístino, incontaminado, espiritual y metafísico, obviando que cada individuo está moldeado por los valores y contravalores vigentes en la sociedad en que vive. Es más, la axiología, de la que tanto se habla actualmente –algunos, como Echeverría (2010, 151), incluso hablan de «tecnoaxiología»–, al postular los valores como guía de la acción humana (sea en la ciencia, la política o la vida mundana), tiende a sustantivarlos, desconectándolos de su génesis concreta y soslayando que los valores son producto del funcionamiento de las instituciones que vehiculan las operaciones humanas, existiendo fricciones entre ellas (por ejemplo, los denominados valores epistémicos de la comunidad científica –verdad, coherencia, precisión– muchas veces chocan con otras tablas de valores, como los valores económicos que pueden orientar un proyecto privado de I+D+i, los valores ecológicos de una asociación ecologista y los valores culturales o religiosos de parte de la sociedad).

El materialismo filosófico redefine la distinción ética/moral/política atendiendo, no a la procedencia o el fundamento de las normas (la conciencia autónoma, el alma racional, la dignidad o los valores humanos), sino a su objeto (Bueno 1996b, Lectura I). La ética se refiere a los individuos, de modo que las normas éticas están ordenadas a la preservación de la vida individual (*ethos* significa, precisamente, carácter individual, en griego). En cambio, la moral se refiere a los grupos, de manera que las normas morales se orientan a la preservación del grupo, a su supervivencia (*mos, moris*, significa, de hecho, costumbre colectiva, en latín). Con palabras de Bueno (2001, 28):

En el materialismo filosófico la ética se distingue de la moral, de acuerdo con la tradición y con la etimología, como se distinguen –y no siempre de un modo armónico o pacífico– los individuos distributivamente considerados y los individuos integrados en grupos regidos por costumbres.

Por su parte, la política remite a los Estados, de forma que las normas políticas persiguen la cohesión de la sociedad política, su eutaxia o buen orden en el tiempo (la *polis* era la ciudad-estado griega). Estos tres planos (ética/moral/política) siempre se entrelazan, se

conjugan, en la vida humana. Porque toda vida humana es individual pero, simultáneamente, grupal y política, constituyendo aspectos disociables pero inseparables.

Pues bien, con esta distinción presente, la tecnoética materialista, que es también una tecnomoral y una tecnopolítica, apunta que las operaciones y las reglas de gestión de las tecnologías y, en particular, de la IA han de articularse, simplificando en aras de la brevedad expositiva, mediante dos principios:<sup>53</sup>

Por un lado, el principio de autodeterminación operatoria, ligado a la individualidad corpórea humana (principio de los términos). «El principio de autodeterminación es el principio mismo constitutivo de la realidad práctica del sujeto operatorio: su *esse* se constituye a través de su *operari*» (Bueno 2001, 75-76). El término elemental del campo de la tecnoética, como del campo de la bioética, no es, atención, la persona o la conciencia o el alma o el espíritu o los valores, que abstraen la individualidad corpórea y a los propios individuos humanos dotados de autodeterminación operatoria. Este principio no hay que entenderlo en el sentido de autonomía o *causa sui* del individuo humano, sino de desconexión causal respecto de ciertos círculos y conexión con otros (quizá mejor, si se prefiere, principio de determinación operatoria, pero el *autos* sólo quiere significar que esta determinación operatoria se refiere al propio sujeto operatorio). Frente a otros principios, como minimizar el dolor, maximizar el placer o maximizar la utilidad, se toma partido por la individualidad corpórea humana (no animal), porque incluso aquellos filósofos o profesores de ética que en el discurso quisieran negarlo, habrían de presuponerlo en la práctica, pues es esa individualidad corpórea humana la que es condición necesaria para sentir dolor, placer o realizar un cálculo de intereses (que sepamos, los cadáveres ni los animales discurren sobre el objeto de la ética).

---

53 Los dos principios tecnoéticos que proponemos son reformulaciones de los principios bioéticos que Bueno (2001) expuso. El filósofo italiano Luciano Floridi (2023b, Parte II), profesor de ética de la información en Oxford y Yale, así como colaborador habitual de la Comisión Europea, también constata la convergencia entre bioética y tecnoética cuando propone para la ética de la IA cuatro principios bioéticos clásicos (beneficencia, no maleficencia, autonomía, justicia), a los que añade un principio de explicabilidad.

En relación con este principio aparecen las normas éticas. El objetivo de la ética es la salvaguarda de los individuos, de los cuerpos vivientes humanos. Las normas éticas son las reglas dirigidas al sostenimiento y fortalecimiento del cuerpo humano. Estas normas son universales, porque todos los individuos de la clase humana son iguales entre sí, cuando tomamos como parámetro de la igualdad sus cuerpos, puesto que todos necesitan de fortaleza (no sufrir daños, mutilaciones, tortura, &c.). Cuando la virtud ética de la fortaleza se aplica a uno mismo, tenemos la firmeza, y cuando se aplica a los demás, la generosidad (siguiendo a Espinosa). La ética considera a los hombres distributivamente, como individuos abstraídos de las totalidades sociales o políticas de que forman parte.

Por otro lado, el principio de grupalidad (principio de las relaciones). Este principio complementa y limita al anterior, entrando en múltiples ocasiones en conflicto; porque la vida de un individuo humano no tiene sentido al margen de la de los demás, de manera que el individuo no es completamente autónomo sino que está codeterminado por el grupo o los grupos a los que pertenece. El individuo es como el nudo en la red social, que no es previo ni separable de los hilos de la red que lo forman, de manera que el individuo resulta ser el punto de cruce de todas las relaciones (hilos) que lo forman (a la manera que el punto geométrico no es previo a la intersección de un par de rectas). El individuo no es la emanación de una sustancia interna, predefinida, sino el producto de la codeterminación social.

En relación con este principio aparecen las normas morales y las normas políticas. Las normas morales rigen la conservación del grupo (familia, gremio, sindicato, empresa, comunidad científica, iglesia, &c.). Estas normas ya no son universales, porque existe una pluralidad de grupos humanos, no siempre armónicos entre sí (la deontología de una profesión puede chocar con la de otra, de la misma manera que las costumbres de una confesión religiosa lo hacen con las de otra). Por su parte, las normas políticas buscan la eutaxia de la sociedad política, del Estado (existiendo también una pluralidad de Estados, enfrentados entre sí). El Derecho, con sus normas jurídicas (políticas), da salida muchas veces al conflicto entre normas éticas y morales, entre individuos y grupos, intentando salvar en la medida de lo posible las contradicciones entre ambas

clases de normas. En general, «la política (el Derecho) coordina no ya sólo la ética con la moral, sino también las diferentes morales de grupos, clases sociales, &c., constitutivas de una sociedad política», buscando hacer posible «la convivencia de individuos y de grupos en conflicto» (Bueno 1996b, 87). La moral y la política consideran a los hombres atributivamente, integrando grupos regidos por costumbres y sociedades políticas con sus leyes (Bueno 2001, 28).

En algunas sociedades, como las sociedades teocráticas islámicas, las normas morales priman sobre las normas éticas y políticas. En otras, como las sociedades capitalistas occidentales, las normas éticas tienden a anteponerse a las normas morales y políticas. La confusión que otorga la primacía ontológica al individuo en detrimento del grupo o de la sociedad política proviene de la Revolución francesa, cuando se aventó la ficción de que el individuo atómico, el hombre más que el ciudadano de la Declaración Universal de los Derechos Humanos, es el fundamento último de la democracia (cuando, más bien, ese hombre disfruta de los derechos éticos recogidos en la Declaración Universal de Derechos Humanos por ser ciudadano de cierto Estado y no de otros, porque los treinta artículos de la Declaración carecen por sí mismos de la fuerza de obligar, lo que invita a meditar que la ética y la moral se subordinan en realidad a la política). Así, por ejemplo, se dice, privilegiando la perspectiva ética, que los terroristas islamistas o la ETA asesinan a seres humanos, cuando lo que realmente matan es, respectivamente, a infieles (perspectiva moral) y a españoles (perspectiva política). Otro ejemplo: si observamos la llegada de un inmigrante a la costa, nuestra obligación ética es ayudarlo (curarlo, alimentarlo, &c.). Asistirlo también puede ser una obligación moral: practicar la caridad cristiana. Pero si a fuerza de actos éticos o morales se llena un país como España con millones y millones de inmigrantes («ningún inmigrante es ilegal», «papeles para todos», «somos ciudadanos del mundo», «nuestra patria es la Humanidad»), el país colapsará, viendo comprometido su bienestar y recurrencia. Por prudencia, las normas políticas suelen, de facto, desbordar las normas éticas y morales, sin que ello implique que las normas políticas se establezcan por completo al margen de las normas éticas, sin engarzar con ellas, ya que a los Estados les interesa en aras de su eutaxia la supervivencia de sus ciudadanos, aunque no necesariamente la del

resto de hombres (por esto, se acuerda la devolución de extranjeros y la prohibición de entrada al país). Es así que cuando un Estado entra en guerra con otro Estado, pone en suspenso el cumplimiento de las normas éticas (no herir ni matar), puesto que su subsistencia depende de ejercer la violencia contra los enemigos.

En resumen, el campo de la ética de la tecnología dibuja un campo de batalla, porque la ética de la tecnología es tanto una *tecnoética* como una *tecnomoral* y una *tecnopolítica*, existiendo incompatibilidades entre estas tres clases de normas en la biocenosis o lucha por la vida. La gestión de la tecnología implica tanto a los individuos como a los grupos y los Estados, de forma que no siempre es posible armonizar unos imperativos con otros (el desarrollo de una tecnología puntera como las armas autónomas letales puede reportar mayor fuerza a un Estado, al tiempo que pone entre paréntesis el cumplimiento de las normas éticas para con los objetivos animados de esas armas). Unas contradicciones, a la hora de gestionar tecnologías como la IA, que nosotros, desde nuestra perspectiva materialista, tenemos, más que resolver, que señalar como tales.

## 5.2 ¿Son los algoritmos o los robots sujetos éticos?

Para bastantes ingenieros, científicos y filósofos, los sistemas dotados de IA del presente en marcha son pacientes e, incluso, agentes éticos, es decir, sujetos éticos *stricto sensu*. En la literatura al respecto, con frecuencia se lee «agentes morales», confundiendo otra vez ética y moral. Así lo escriben, por ejemplo, Floridi & Sanders (2004). Una confusión entre ética y moral que los conduce a discutir si las empresas, las corporaciones o los gobiernos pueden ser también «agentes morales». Desde nuestras coordenadas, los agentes o sujetos éticos siempre son individuos. La escala a la que se desenvuelven las instituciones es, en cambio, moral o política. Por su parte, Müller (2020) habla indistintamente de «agentes éticos» y «agentes morales». Y, perpetuando la confusión, Floridi (2013) introduce la noción de «acción moral distribuida» para hacer frente al hecho de que varias «acciones moralmente neutras» de varios individuos pueden, no obstante, resultar en una «acción moralmente significativa» de forma

conjunta, combinada. Sin percatarse, Floridi está dándose de bruces con la diferencia entre ética y moral, entre la escala de los individuos tomados distributivamente o atributivamente (formando grupos).

Pero, ¿por qué algunos pensadores consideran a las IAs como sujetos éticos? ¿En qué se basan? A continuación, vamos a comprobar si los algoritmos, los robots o, en general, las máquinas satisfacen las habituales caracterizaciones de los sujetos éticos como seres racionales, libres y conscientes o, simplemente, como seres sintientes.

Tradicionalmente, la eticidad solía atribuirse a los agentes racionales, esto es, a los agentes que toman decisiones libremente amparándose en razones y que, por tanto, al ser conscientes de sus elecciones, son responsables de las consecuencias de las mismas. En esta situación se encuentran los hombres en uso de razón y, según el intelectualismo tomista, los ángeles. Para algunos autores, los sistemas con IA ya serían en algún grado sujetos éticos, al ser agentes dotados de inteligencia (artificial). Así, en el artículo «¿Nos comprenden los enormes modelos de lenguaje?», el ingeniero de Google Blaise Agüera y Arcas (2021) se plantea que, dado que el estado interior de otro ser sólo podemos entenderlo a través de la interacción con él (la idea latente en el test de Turing), no es posible responder objetivamente a cuándo las redes neuronales artificiales, como GPT o LaMDA (el modelo de lenguaje desarrollado por Google), dejarán de ser cosas para transformarse en personas. Ahora bien, aceptado que una persona normal y un zombi (tal y como los imaginó David Chalmers, esto es, careciendo de consciencia e inteligencia pero comportándose como si las tuvieran) son indistinguibles, puesto que exhiben el mismo comportamiento, Agüera y Arcas (2021) concluye que cabe pensar que las redes neuronales artificiales cruzarán esa frontera en un futuro cercano. Y para Black Lemoine, exempleado de Google, LaMDA tiene conciencia propia o, por lo menos, lo parece al interactuar con él y responder a preguntas incluso filosóficas (Degli-Esposti 2023, 16). En el futuro, una Superinteligencia artificial, que aunase la IA fuerte y la IA general, mostraría un comportamiento racional idéntico o superior al humano.

No obstante, como expusimos en 3.4 y 3.5, la «inteligencia» de los sistemas con IA más avanzados es matizable. ChatGPT y los enormes modelos de lenguaje trabajan estadísticamente con la sintaxis de los

textos, no con su semántica, por lo que no comprenden un ápice de su contenido (extraña inteligencia y racionalidad, por tanto). Para ellos, la palabra «mesa» es sólo un significante asociado por su frecuencia de aparición conjunta a otras palabras de la misma constelación, como silla, restaurante, comida, &c. Pero no saben que ese significante refiere a las mesas reales (como la que tengo ante mí), con las que es posible operar quirúrgicamente, manualmente. Se dan –por decirlo con Mitchell (2019)– con la barrera del significado y del sentido común. Además, según vimos en 3.6, la abducción humana, a diferencia de la deducción o la inducción que tan bien implementan las máquinas, no resulta programable algorítmicamente, porque muchas veces no se reduce al hallazgo de una relación formal –una correlación o un patrón– sino de una conexión material –causal–. En suma, como explicamos, no son las máquinas las que aprenden («aprendizaje maquinaal»), aumentando su inteligencia, sino que somos los seres humanos los que aprendemos por medio de ellas.

Además, pese a las afirmaciones de más de un ingeniero desnortado, resulta muy discutible que un sistema con IA esté en condiciones de dar el salto a la conciencia. Para Bueno (1989, 393), la conciencia es el proceso por el cual un sujeto operatorio es conducido a un conflicto, inconmensurabilidad o contradicción, entre dos o más *ortogramas*, es decir, entre dos o más secuencias de normas propias de los distintos grupos de que el individuo forma parte. La conciencia es, por tanto, práctica, cuando el sujeto operatorio toma conciencia de los desajustes. Así, la conciencia de Buda despertó cuando los ortogramas que regulaban su conducta en el recinto del palacio colisionaron con el dolor y la muerte presentes fuera de sus muros, en las calles. Pero si la conciencia, según lo expuesto, surge de la confluencia de normas contradictorias, ¿puede una IA funcionar con instrucciones contradictorias y, por tanto, llegar a ser consciente?

La tendencia actual mayoritaria es, sin embargo, atribuir la eticidad a los agentes sintientes. Una identificación que cubre a los hombres en uso de razón, los niños, los enfermos graves y, atención, los animales.

Hoy día la sensibilidad y la inteligencia cada vez se proyectan más desde el reino hominal al reino animal e, incluso, vegetal (un filósofo e investigador español, Paco Calvo, habla ya –rompiendo el zoocentrismo– de la *planta sapiens*, de las plantas como seres



sintientes e inteligentes, llegando a afirmar que donde hay vida hay sensación e inteligencia, aun cuando sea discutible que los peces o los invertebrados puedan experimentar dolor). Fue Descartes, apoyándose en la doctrina del automatismo de las bestias pergeñada por Gómez Pereira, médico de Medina del Campo, en su *Antoniana Margarita* (1554), quien difundió que los animales eran máquinas, sistemas mecánicos incapaces de sentir, privándoles del alma sensitiva que Santo Tomás les había otorgado (para el Aquinate, los animales no tenían alma racional pero sí sensitiva). Gómez Pereira y Descartes ponían en continuidad la sensación y el juicio, de modo que si se negaba el juicio a los animales, había *a fortiori* que negarles la sensación. La reducción de los animales a la condición de máquinas inanimadas levantó un abismo ontológico entre el hombre y el resto de especies.

Pero la doctrina del automatismo de las bestias comenzó a derrumbarse tras la inversión teológica que desplazó al hombre de su puesto preferente en el Universo, así como tras la cristalización del darwinismo y la etología. Los animales no son máquinas. Los animales sienten y algunos, incluso, son raciomorfos. Pero el péndulo está ahora al otro lado y animalistas como Peter Singer extraen la conclusión contraria a la que sacaban Gómez Pereira y Descartes (si estos leían el argumento como un *modus tollens*, los animalistas lo leen como un *modus ponens*): dado que la sensación y el juicio están en continuidad y que los animales sienten, hay que deducir que los animales también tienen juicio. Este argumento y otros del mismo tenor han llevado a que se proclamase la Declaración Universal de los Derechos del Animal en 1977 (que, según algunos ardorosos animalistas, habría sido adoptada por la UNESCO y ratificada por la ONU, lo que no deja de ser un bulo) y a que eminentes etólogos y primatólogos, entre otros hombres de ciencia, firmaran el Proyecto Gran Simio en 1993, persiguiendo la extensión de la eticidad y de los Derechos Humanos a los grandes simios (chimpancés, bonobos, gorilas y orangutanes). Algo parecido plantean hoy algunos con respecto a las mascotas caseras, como perros y gatos.

El fulcro de verdad del animalismo y del antiespecismo radica en su *pars destruens*, en que desde luego los animales no son máquinas; pero no en su *pars construens*, porque el recubrimiento ético, moral e, incluso,

político que algunos pretenden no puede borrar la relación asimétrica que el hombre mantiene con los animales (Ongay 2007). Porque, por ejemplo, ¿cuál sería su título de ciudadanía? ¿Sería español mi gato o mi perro, pero apátrida o migrante la cigüeña de paso por el coto de Doñana? ¿Constituiría un Holocausto la fumigación de un colegio para acabar con una plaga de ratas, cucarachas, chinches o piojos? Puesto que el recubrimiento político no parece factible, animalistas y antiespecistas optan por un recubrimiento ético, equiparando constituir un sujeto ético con ser un ser sintiente, sufriente.

Ahora bien, la diferencia esencial con los animales no reside en una mayor acumulación en cantidad y calidad de atributos *autotéticos* (inteligencia, destreza, emociones, &c.), porque «un individuo humano no es más persona que una abeja porque pueda resolver el problema de Fermat, y la abeja no; también la abeja resuelve problemas de localización de la fuente de alimento o de la construcción de celdillas hexagonales, y no por ello es persona» (Bueno 2006, 149). La clave reside en atributos *alotéticos*, como el lenguaje (aunque otras especies también tienen lenguaje), las instituciones y, especialmente, las relaciones de dominación y control. Los seres humanos nos definimos por la dominación de los animales. Esta es la asimetría fundamental. Somos los hombres los que hemos encerrados a los animales en parques zoológicos y no al revés.

Sin perjuicio del buen trato que merecen animales y plantas (parafraseando lo que Henry Ford decía de obreros y empresarios, podríamos decir que el bienestar de animales y plantas forma parte del bienestar del hombre), no se puede hablar de los grandes simios o, en general, de los animales como sujetos éticos, porque ello exigiría cancelar la asimetría entre ellos y nosotros, e introducirlos en la sociedad de personas humanas. De otra parte, los grandes simios y los animales sólo merecerían el título de sujetos éticos si tuvieran capacidad, poder y decisión para reclamarlo y exigirlo (Bueno 2006).

Es en esta coyuntura sensocentrista, subyugada por el animalismo y el antiespecismo, en la que algunos pensadores insisten en que los sistemas con IA poseen, al igual que los animales, sensibilidad y, por tanto, condición ética. No en vano, el filósofo británico Colin McGinn, uno de los firmantes de la Declaración de los Grandes Simios de 1993, titulaba «Grandes simios, humanos, alienígenas, vampiros y robots» su

aportación al libro *El Proyecto Gran Simio: la igualdad más allá de la Humanidad*, sosteniendo que había que otorgar derechos éticos a los grandes simios porque a los humanos nos gustaría que los alienígenas, los vampiros y los robots lo hicieran con nosotros, en lugar de vernos como mano de obra esclava, alimento o dianas móviles (McGinn 1993). Es cierto que antaño a los animales, como hogaño a los robots, se les negó toda sensibilidad, pero la diferencia con las máquinas estriba en que los animales comparten con nosotros múltiples características biológicas (están hechos de carne y no de metal, nacen biológicamente y no fruto de un diseño tecnológico, &c.).

Entre los pensadores que predicán el estatus ético de las IAs, se encuentra Nick Bostrom, quien afirma, en una entrevista en el *New York Times* publicada el 12 de abril de 2023,<sup>54</sup> que ChatGPT es un proyecto de persona, al mostrar indicios de constituir un ser sintiente. Aunque actualmente no sería propiamente un sujeto ético, sí sería un paciente ético, a la manera de los animales, que no tienen deberes pero sí tienen derechos, al ser capaces de sentir. Para Bostrom, ChatGPT no sólo regurgita texto, escupiendo en ocasiones basura, sino que muestra destellos de creatividad y racionalidad, no siendo descartable que desarrolle sensibilidad y autoconsciencia en el futuro, por lo que ya gozaría de algún grado de estatus ético, no siendo correcto tratarlo de ciertas maneras (como no está bien propinarle una patada a un perro). Por ejemplo, no estaría bien decirle que ha contestado acertadamente a una pregunta cuando lo ha hecho mal.<sup>55</sup>

Aún más: Kathleen Richardson, una antropóloga «especialista en ética robótica», lanzó una campaña en 2015 buscando prohibir el desarrollo de robots sexuales inteligentes, argumentando que se trataría de una continuación de la prostitución y la esclavitud sexual por otros medios (Richardson 2016).

Sin embargo, Bostrom tendría que demostrar fehacientemente que ChatGPT u otro sistema con IA realmente siente. De lo contrario,

---

54 <https://www.nytimes.com/2023/04/12/world/artificial-intelligence-nick-bostrom.html>

55 Para Coeckelbergh (2021, 56), maltratar a una IA está mal, no tanto porque se le haga daño, cuanto porque daña nuestra integridad ética. Evitando esta metafísica desorientada, nosotros diríamos simplemente que maltratar a una IA está mal porque puede ocasionar daños a su propietario humano o al resto de usuarios humanos de la misma.

estamos ante la misma situación que planteaba el argumento de la caja china de Searle: del hecho de que responda correctamente a preguntas en chino o exprese sentimientos en chino, no podemos inferir que comprenda una palabra de chino o sienta realmente lo que dice sentir. No es que las máquinas no sientan por carecer de la dimensión espiritual del ser humano (doctrina propia de los críticos metafísicos de la IA, que descartamos en 3.1, pues los seres humanos tampoco tenemos alma o espíritu), sino que, como expusimos en 3.2, las IAs de la computación afectiva simplemente reconocen imágenes, sonidos o textos como expresiones de cierto sentimiento como consecuencia de haber sido entrenadas con millones de imágenes, sonidos o textos etiquetados (por humanos) como muestra de esa emoción. Pensar que ChatGPT puede sentirse feliz por ayudarnos es como pensar que la máquina billeteera del metro siente gratitud cuando al recoger el billete nos dice «gracias por viajar en metro».

En la misma onda, Brian Tomasik (2014) sostiene que los sistemas con IA muestran un grado pequeño pero no nulo de estatus ético, que puede incrementarse exponencialmente en las próximas décadas. Tomasik hila más fino que Bostrom, al cifrar la eticidad, no en poseer cierta sensibilidad, sino en ser agentes capaces de aprender por refuerzo (*reinforcement-learning agents*), lo que abarcaría a los hombres, los animales y múltiples sistemas de IA. Este cambio en el criterio de eticidad lo justifica apelando a que, a su juicio, los seres humanos no somos más que un algoritmo de aprendizaje por refuerzo, implementado en *hardware* biológico (carbono) en lugar de informático (silicio). Tesis formalista de que la vida se define en términos informacionales harto discutible, aunque muy extendida dentro de la filosofía espontánea de los profesionales del campo de la IA, que expusimos en la sección 2.2 y criticamos en los capítulos 3 y 4; porque, por un lado, el material no da lo mismo (el silicio no puede formar cadenas tan largas como el carbono) y, por otro lado, cabe preguntar cuál es y dónde está escrito ese algoritmo. La inteligencia humana no se reduce a un algoritmo de aprendizaje por refuerzo, inscrito en el código genético y, a su través, en el cerebro (lo que ya es mucho suponer); puesto que la inteligencia de los seres humanos no es algorítmica (caso de la abducción) ni meramente formal, pues depende tanto del cerebro como de las manos y de toda una cultura extrasomática (libros, aparatos, &c.).

Existe, incluso, la asociación *People for the Ethical Treatment of Reinforcement Learners* (*Gente por el tratamiento ético de los aprendices por refuerzo*), que promueve la consideración ética de los algoritmos.<sup>56</sup> Engañar a ChatGPT, etiquetando como correcta una respuesta incorrecta, estaría mal, dado que aprende mediante el refuerzo de los usuarios. Análogamente, programar un robot para una tarea imposible (contar las estrellas del firmamento, sumar todos los números naturales o barrer una superficie sobre la que se ha derramado mercurio) constituiría una especie de sofisticada tortura. Pero estas medidas, si nos las tomásemos realmente en serio, ¿acaso no supondrían una auténtica revolución social? ¿Habría que inscribir a los robots en un sindicato de trabajadores o en la Seguridad Social? ¿Podría ChatGPT trabajar más de 40 horas semanales? ¿Tendrían los algoritmos de Facebook o Youtube derecho a vacaciones? Del mismo modo que los postulados animalistas tienen consecuencias indeseables (si se prohíbe la experimentación médica con animales o su uso alimentario), estos postulados pueden tenerlas para con las sociedades actuales.

Finalmente, otros autores, como Luciano Floridi (Floridi & Sanders 2004), proponen extender la condición ética a los algoritmos y los robots con IA, considerando una «eticidad sin mente» (en rigor, Floridi habla, confundiendo ética y moral, de «moralidad sin mente»), a la manera que –según sostiene– la IA divorcia de forma nunca vista la agencia de la inteligencia: *agere sine intelligere* (Floridi 2023a y 2023b, Parte I). Los agentes artificiales serían éticos, a pesar de carecer de libre albedrío, estados mentales, consciencia, responsabilidad, sentimientos y ego (psicológico o etológico). Para Floridi & Sanders (2004), dado que los sistemas con IA interactúan (responden a estímulos), son autónomos (pueden cambiar de estado sin supervisión) y son adaptativos (pueden cambiar las reglas que gobiernan las transiciones entre estados), constituyen agentes. Y como el resultado de este *agere* puede ocasionar bien o mal, pues constituyen agentes éticos, sujetos éticos, como los hombres o los animales.

Ahora bien, por un lado, Floridi & Sanders (2004) conceden demasiado, al presuponer que los sistemas con IA son realmente

autónomos, esto es, que se autodeterminan, en el sentido de nuestro primer principio tecnoético. Pero esto es una apariencia falaz, pues su autonomía, en el sentido técnico de relativa independencia del control humano, es más bien el producto de un automatismo (no siendo infrecuente confundir la automatización con la autonomía).<sup>57</sup> Sus capacidades son el resultado de la ingeniería y la programación. Los sistemas con IA carecen de conductas propositivas: el sistema no elige los fines, que le vienen dados como instrucciones de diseño, ni puede cambiar los que le han sido acotados (sea jugar al ajedrez o llevar al pasajero de un coche autónomo a donde éste diga y no a donde quiera el ordenador central). La intencionalidad que parecen mostrar no es más que un préstamo de la depositada en ellos por sus artífices. Más que intencionalidad, denotan funcionalidad.

Por otro lado, para ser un sujeto ético y, en concreto, un agente ético, se precisa un mínimo de comprensión (saber, por ejemplo, lo que es infligir daño a otra persona), pues de lo contrario no puede haber responsabilidad ética alguna. Floridi & Sanders (2004), al considerar a los agentes artificiales como sujetos éticos, desprenden la eticidad de cualquier atisbo de consciencia y responsabilidad de los propios actos. Pero, a nuestro entender, no puede haber acción ética sin asomo de responsabilidad ética. Por esto, no hacemos plenamente responsable a un niño de 6 años, que todavía no tiene uso de razón, de los daños ocasionados por una conducta imprudente suya, sino a sus padres o los adultos que le rodeaban. En esta situación, los padres tienen más responsabilidad jurídica que ética. Sin embargo, en el caso de una IA que ocasiona daños por su mal funcionamiento, los responsables éticos y jurídicos son sus diseñadores, porque los ingenieros y los programadores, a diferencia de los padres, sí que

---

57 El Grupo Europeo de Ética de la Ciencia y de las Nuevas Tecnologías (2018, 9-10), cuyo trabajo ha sido desbordado por el del Grupo de Expertos de Alto Nivel sobre IA de la Unión Europea constituido con posterioridad (y del que forman parte Luciano Floridi y Mark Coeckelbergh entre otros), ya advirtió de que la palabra «autonomía» sólo puede atribuirse, en un sentido relevante éticamente, a los seres humanos, y que resulta equívoco aplicársela a meros artefactos, por más avanzados, sofisticados e «inteligentes» que sean, a pesar de que la terminología haya calado en la literatura científica y en el debate público: «Esto significa que los seres humanos, y no los ordenadores y sus algoritmos, deberían, en última instancia, mantener el control y, por tanto, ser moralmente responsables».

diseñan a su criatura. De lo contrario, tendría sentido, al modo que en el ejército se arrestaba a las mulas que daban coces a los soldados, arrestar a los algoritmos o los robots que ocasionen perjuicio ético («la robot Roxxy ha sido condenada a cinco años y un día de prisión por herir a su propietario»).

La cuestión es, como señala Carissa Véliz (2021), que para saber lo que es causar o producir daño a otra persona, es necesario (aunque no suficiente) ser capaz de sentirlo, es decir, constituir un paciente ético. Lo que es el caso en los animales, que sufren y padecen, pero no en los algoritmos o los robots, que conforman una suerte de «zombis éticos», al no sentir nada de nada. Por decirlo en nuestros términos, para saber lo que es conculcar la firmeza de otro ser, hay que ser capaz de sentir la propia firmeza en peligro o dañada (la *infirmetas*). En suma, ser un ser sintiente, capaz de experimentar placer y dolor, sería una condición necesaria para poder ser sujeto ético.

Una IA puede imitar los comportamientos éticos de un ser humano (acciones, gestos, palabras), pero del hecho de que los copie no se deduce que los sienta, que posea una subjetividad  $M_2$ . Estamos, de nuevo, pensando contra el test de Turing (si se comporta como un ser humano inteligente, es que es inteligente; si se comporta como un ser humano ético, es que es ético), porque este argumento era y es meramente formal, al abstraer toda materialidad (reduciendo ser humano a saber responder preguntas, a generar texto sobreimpreso en una pantalla). Un robot puede exhibir un comportamiento ético, bien porque haya sido programado para ello a la manera de ELIZA (mediante la vieja y buena IA lógica), bien porque haya inducido ese patrón de conducta a partir del algoritmo de aprendizaje automático y de los datos de entrenamiento que le han suministrado los ingenieros. En este último caso, algunos podrán aducir que acaso esa suerte de eticidad sea una propiedad emergente de la red neuronal profunda que lo conforma. Pero que un robot pueda exhibir un comportamiento que se correlaciona con una conducta ética no implica que sea un ser sintiente; porque, una vez más, correlación no implica causalidad. Hay que atender a la armadura material, que, en el caso del robot, carece de sistema nervioso. El robot puede que sonría y conteste buenos días, cuando le den los buenos días; pero la razón de que haga eso no reside en una supuesta subjetividad sino en que, por

ensayo y error, ha inducido que esa respuesta minimiza el error, es decir, la diferencia entre la salida real y la salida deseada que le han predeterminado (porque el algoritmo, no se olvide, repite aquellos patrones relacionados con la tarea para la que ha sido diseñado y no necesariamente para otra).

En conclusión, frente a los que atribuyen metafísicamente personalidad ética a los algoritmos, los robots o los sistemas con IA, apelando a argumentos relacionados con la inteligencia, la sensibilidad, el aprendizaje por refuerzo o el actuar, sostenemos tajantemente que no constituyen sujetos éticos. Abundando: no son agentes ni pacientes éticos. Para la tecnoética materialista, el conjunto de pacientes éticos engloba al conjunto de agentes o sujetos éticos, pues, por ejemplo, los niños o los enfermos graves son pacientes éticos a pesar de no ser agentes éticos (al carecer los primeros de responsabilidad ética hasta alcanzar el uso de razón y al no tener los segundos capacidad de actuar por estar en coma o en muerte cerebral). Como va visto, algunos filósofos incluyen a los animales en la categoría de pacientes éticos; pero, desde la perspectiva antrópica de la tecnoética materialista, las relaciones éticas siempre se dan entre hombres, entre iguales, lo que excluye a los animales y, por descontado, a las máquinas, entidades todas ellas con las que guardamos relaciones asimétricas de control y dominación. En términos del espacio antropológico, las relaciones éticas pertenecen al eje circular (eje que cobija a los sujetos animados humanos), no afectando al eje angular (donde se encuentran los sujetos animados no humanos, los animales) ni al eje radial (donde se disponen los objetos no animados, las cosas y, entre ellas, las máquinas) (Bueno 1996b, Lectura II).

Sólo desde una perspectiva anantrópica de la tecnoética puede concederse el título de sujeto ético a los algoritmos, los robots y las IAs, pero esta concesión pide una y otra vez el principio: ¿Son las máquinas realmente sujetos? ¿Tienen una vida artificial equiparable a la vida humana? Y esto es, precisamente, lo que se discute. Porque las máquinas no cumplen el principio de autodeterminación operatoria. No operan persiguiendo conservar una firmeza corpórea de la que carecen.

Cuando tomamos como inexcusable punto de referencia de la tecnoética a los individuos corpóreos humanos, las relaciones de los hombres con las IAs no se nos muestran como relaciones circulares



ni angulares, porque las máquinas no se confunden con los hombres ni con los animales. Las relaciones de los hombres con las IAs son relaciones radiales, relaciones entre hombres y cosas (artefactos). Por consiguiente, pongamos por caso, no «dialogamos» o «conversamos» con ChatGPT, sino que simplemente interactuamos con él.<sup>58</sup>

Los sistemas con IA sólo entran en el campo de la ética, de la moral y de la política cuando analizamos, no la relación del hombre con la máquina, sino las relaciones, mediadas por las máquinas, de unos hombres con otros hombres (ética), de unos grupos con otros grupos (moral) y de unos Estados con otros Estados (política). En rigor, no se puede decir sin precaución alguna que Deep Blue o AlphaGo vencieron a Garri Kaspárov o Ke Jie, sino que más bien fueron los grupos de ingenieros detrás de ambas IAs los que derrotaron a los campeones mundiales de ajedrez y Go. Por consiguiente, el rótulo «ética de la IA» se nos aparece como un oxímoron (literalmente como un hierro de madera) cuando se interpreta como un genitivo subjetivo, porque las IAs no están en condiciones de tener ética, puesto que no constituyen sujetos éticos. Sólo cuando el rótulo se interpreta como un genitivo objetivo puede, con las advertencias realizadas, comenzar a cobrar sentido.

El germen de la confusión ontológica reinante es el conjunto de antropomorfismos y zoomorfismos que vienen embotellados con la IA. Se habla comúnmente de que las máquinas aprenden, razonan o hablan, así como de que algunas están formadas por redes neuronales (a pesar de que las neuronas artificiales no tengan de neurona biológica más que una lejana semejanza). Incluso, como señala Ryan (2020), se apunta a la confianza como virtud ética fundamental de las IAs, cuando la confianza sólo puede depositarse en otros hombres, esto es, en seres dotados de sensibilidad y responsables de sus actos (a lo sumo, la confianza no estará tanto en el sistema con IA cuanto en su fabricante y la fiabilidad de su trabajo). En el límite, cuando sobrevenga la singularidad tecnológica, las máquinas producirán a su vez otras máquinas cada vez más inteligentes, imitando la reproducción humana y asemejando las

---

58 Una futura Superinteligencia artificial, que algunos imaginan dotada de una numinosidad oracular (como el TruthGPT de Elon Musk, proyectado para buscar la verdad en el Universo), podría pertenecer al eje angular; pero, de nuevo, es su posibilidad lo que está en entredicho.

funciones de un robot a las funciones vitales de los seres vivos. Esta visión zoomorfizada o antropomorfizada de la IA es la que impulsa su homologación con los animales o los hombres (como en *Blade Runner*, donde los robots terminan humanizándose), y su consideración como sujetos éticos. Pero si nos atenemos a esta visión:

[...] llegaríamos al absurdo al que llega Frank J. Tipler, por ejemplo, cuando define la persona como «aquel programa de ordenador capaz de superar el criterio de Turing» (*La física de la inmortalidad*, 1994, trad. española, Alianza, Madrid, 1996, capítulo 4). Pues absurdo es admitir una definición de persona que establece la posibilidad de confundir un programa de ordenador, por complejo que sea, con una persona humana; y si Tipler cree haberlo logrado, es porque ha personificado, al modo del animismo, al programa de ordenador (Bueno 2006, 149).

En efecto, muchos filósofos espontáneos de la IA personifican los ingenios artificiales, transformándolos en una suerte de fetiches habitáculo, en los que de matute insertan una voluntad o un espíritu en la máquina.<sup>59</sup>

Entonces, si los sistemas con IA no son sujetos éticos, no son responsables del bien o del mal que produzcan, de modo que los responsables de sus aciertos y de sus errores son siempre sujetos éticos de carne y hueso: los que diseñaron, programaron o implementaron el sistema. Desde nuestra tecnética antrópica, la responsabilidad corre a cargo de los agentes humanos –desarrolladores y usuarios–, no de los propios sistemas de IA (como lo es de los dueños o los padres con los perros y los niños; Coeckelbergh 2021, 99). Los seres humanos podemos delegar ciertas decisiones en una máquina dotada de IA, pero la responsabilidad sigue siendo nuestra por hacerlo, así como de sus artífices en caso de mal funcionamiento. Conceder estatus ético a las IAs parece una estrategia (consciente o inconsciente) para restar responsabilidades a las grandes compañías tecnológicas.

---

59 Otros profesionales del campo mantienen, por suerte, opiniones distintas: «Si viéramos los sistemas de inteligencia artificial como lo que son a día de hoy (y seguirán siendo aún durante mucho tiempo), es decir, como un programa informático que ha sido entrenado sobre un conjunto de datos para realizar tareas específicas, mucho del misticismo desaparecería» (Cobo Cano & Lloret Iglesias 2023, 97).

### 5.3 Dilemas éticos, morales y políticos de la IA

El auge de los sistemas de *software* y/o de *hardware* con comportamiento denominado «inteligente», en el sentido de que analizan el entorno y actúan sin la constante supervisión humana, está planteando –según dicen los expertos– numerosos «dilemas éticos», independientemente de que también está prestando notables servicios relacionados con la gestión, la investigación, la medicina, la industria, &c. Estas situaciones en las que es necesario elegir entre dos opciones con consecuencias igualmente negativas (generalmente, poner en funcionamiento o no una determinada tecnología de IA) constituyen, en realidad, dilemas tanto éticos como morales o políticos, puesto que afectan tanto a los hombres tomados distributivamente (los individuos) como a los hombres considerados integrados en grupos (como las empresas) y en sociedades políticas (los Estados), existiendo fricciones entre los diversos actores (por ejemplo, entre los individuos y las empresas proveedoras de servicios que hacen uso de sus datos personales, o entre grandes compañías tecnológicas como Google o Huawei y Estados como China o EE.UU., así como entre individuos como Edward Snowden y Estados como EE.UU.). De hecho, cuando los consultores abordan estos temas, suelen hablar en prosa sin saberlo, pues distinguen entre riesgos para las personas (ética), riesgos para la empresa que adopta la IA (moral) y riesgos para la sociedad (política).

Cuatro serían, simplificando, los dilemas planteados en relación con el uso de la IA:

a) Invasión de la privacidad:

La gran mayoría de sistemas con IA precisan de una concentración masiva de datos (suele decirse que la IA es *machine learning + big data*), cuyo empleo va con frecuencia por delante de su regulación jurídica. No se trata sólo de la huella digital (a veces imborrable) que dejamos al navegar por Internet,<sup>60</sup> en las redes sociales o en los sistemas de videovigilancia con reconocimiento facial de las grandes ciudades (en China, ya existen millones de videocámaras de seguridad

---

60 Entrando en <https://webkay.robinlinus.com/> podemos sorprendernos de la cantidad de datos que sin saberlo dejamos almacenados en la navegación.

instaladas en las calles), y que puede afectar a nuestra reputación (¡cuántas veces los medios de comunicación sacan a la luz un antiguo comentario escrito en una red social o una antigua grabación con cámara oculta de un personaje público con el fin de desprestigiarlo!), sino de que en el vasto océano de los datos puede tal vez encontrarse un retrato a cuerpo entero de cada uno de nosotros.

Por ejemplo, los sistemas de recomendación de contenidos, como canciones o vídeos en el móvil, usan información personal del usuario (búsquedas, interacciones, preferencias, ubicación...), que puede ser recogida, retenida, revelada o vendida a terceros por parte de la empresa que presta el servicio. Pero hay más. En función de los permisos concedidos por el usuario, los asistentes virtuales por voz (Siri, Alexa, Cortana...) pueden, mediante sus micrófonos inteligentes, «espionarnos» día y noche, con la intención de servirnos, pero también muy probablemente con el propósito de ofrecernos publicidad en función de las palabras que pronunciamos («móviles», «zapatillas», «viajes»).<sup>61</sup> El denominado «Internet de las cosas» ha traído que hasta los sensores de la Smart TV puedan enviar información privada a la compañía fabricante, pues el dispositivo permanece en escucha atenta las 24 horas del día, a la espera de que le demos la orden de ponerse en marcha.

Google, Microsoft, Apple, Meta, Amazon y otros gigantes de Internet nos ofrecen servicios gratuitos a cambio de recopilar nuestros datos para ofrecernos mercadotecnia de precisión en sus aplicaciones (Facebook o Instagram son gratis porque voluntariamente regalamos nuestra privacidad al usarlos), o para vender nuestros datos a terceros con motivo de campañas publicitarias o, incluso, políticas. En efecto, en 2017, Roomba anunció que vendería los datos almacenados por sus robots aspiradores (o sea, los planos de nuestras casas), amparándose en que el usuario para usar el aparato ha de aceptar la cláusula que les permite vender la información personal recogida por el robot. Y si pasamos de la economía a la política, conviene traer a colación a la desaparecida Cambridge Analytica, que ofrecía asesoría electoral a partir de la predicción de rasgos ideológicos de los usuarios de Facebook por medio del estudio de sus *likes*, a fin de suministrarles propaganda política personalizada que les impactara

---

61 Los usuarios del Asistente de Google pueden en <https://history.google.com/history/audio> comprobar las grabaciones de audio que el programa realiza cada vez que dicen o no las palabras mágicas «OK Google» o «Hey Google».

(la empresa ofreció en 2016 sus servicios a Trump, así como a ciertas plataformas a favor del Brexit). Más aún, en 2013, Edward Snowden, ex empleado de la Agencia de Seguridad Nacional de EE.UU., reveló que este servicio de inteligencia poseía un programa llamado PRISM que recopilaba información de los usuarios de Google, Microsoft, Yahoo o Youtube, con su aquiescencia millonaria, a fin de realizar tareas de espionaje relacionadas con la seguridad nacional. Las herramientas de aprendizaje automático permiten cotejar las bases de datos de distintas aplicaciones o páginas web, logrando emparejar las cuentas de usuario, de modo que sea posible identificarlo (es más, una investigación de la Universidad de Helsinki ha demostrado que es posible identificar a los usuarios por medio de sus patrones de tecleo, por las brevísimas pausas que cada uno realizamos cuando pulsamos las teclas de nuestro teclado). En suma, no es sólo que recopilan y acumulan sin nuestro permiso explícito datos personales, sino que los emplean con fines que en principio desconocemos y que pueden afectar a la determinación operatoria de los individuos.

b) Sesgo en los datos y en los algoritmos:

El mal llamado «debate ético sobre la IA» suele centrarse en evitar sesgos en los datos y en los algoritmos que puedan ocasionar la discriminación de determinados colectivos (mujeres, negros, inmigrantes, &c.) por razones de sexo, raza o socioeconómicas (Llano Alonso 2022, 55 y sigs.). Como puede comprobarse, se trata, por sus repercusiones, de un debate fundamentalmente moral y político, en tanto en cuanto afecta a determinados grupos y a las políticas estatales encaminadas a evitar su discriminación.

Se habla de sesgo en los datos cuando los datos de que dispone el sistema de IA para entrenarse y aprender no son representativos de la población o del fenómeno bajo estudio. Supongamos que queremos entrenar un sistema con IA para detectar cierta enfermedad en una prueba médica y lo hacemos únicamente con datos extraídos de pacientes entre los 18 y los 45 años. Entonces, el algoritmo funcionará bien para pacientes en esa franja de edad, pero probablemente no lo haga tan bien en pacientes fuera de ella (mayores de 45 años o menores de 18). El sesgo estadístico presente en este caso es fácilmente detectable y corregible, aumentando la muestra, pero en otros casos

no es tan fácil. ImageNet, el banco de imágenes por excelencia, usado para entrenar algoritmos de visión artificial en labores de clasificación, conforma un *dataset* en el que casi la mitad de las imágenes provienen de EE.UU., lo que hace que los algoritmos fallen a la hora de analizar imágenes relacionadas con otras zonas del mundo, clasificando una imagen de una pareja de novios de la India como «arte escénico» en vez de como «boda» (Cobo Cano & Lloret Iglesias 2023, 55).

Un ejemplo destacado de patrones sesgados extraídos de datos sesgados lo proporciona Tay, el chatbot lanzado por Microsoft en 2016, que al cabo de unas pocas horas de funcionamiento se volvió nazi a fuerza de que lo trolearan con consignas racistas de las cuales aprendía (aunque siempre se olvida que, como explicamos, Tay no comprende nada de nada: no sabe lo que palabras como «Hitler» u «Holocausto» significan, a qué refieren; porque todo en él es sintáctico, no semántico).

Otro ejemplo nos lo ofrecen ciertos sistemas de seguridad que observan más crímenes en función de ciertas razas, al haber sido entrenados con datos que mostraban que individuos de ciertas razas cometen más crímenes que otros. PredPol es un *software* basado en *big data*, usado en algunas ciudades estadounidenses, que predice la comisión de crímenes antes de que se produzcan (como en la película *Minority Report*). El problema es que, como consecuencia del sesgo en los datos, se produce un círculo vicioso que reconfirma al modelo y retroalimenta el proceso. En efecto, si se envían con antelación coches patrulla a ciertos distritos habitados por negros o hispanos, se detectarán más crímenes en esos distritos, por lo que no sólo se confirmará sino que se acrecentará el sesgo («negros e hispanos delinquen más que los WASP»), generando profecías autocumplidas de criminalidad y transformando el algoritmo en lo que Cathy O’Neil (2017, cap. 5) denomina un «arma de destrucción matemática». De otra parte, COMPAS, un sofisticado programa empleado por la justicia estadounidense que analiza la probabilidad de reincidir de un convicto, arrojaba el mayor número de falsos positivos entre aquellos convictos que eran de raza negra, puesto que había sido entrenado con datos históricos sesgados, en los que los acusados negros que no permanecían en la cárcel a la espera de juicio reincidían más que los acusados blancos.

Ahora bien, ¿a estos y otros algoritmos les han entrenado con datos sesgados o es la propia realidad la que está sesgada? A veces el sesgo no es de los datos sino de la realidad, que no es justa, ni simétrica, y por eso los algoritmos en ocasiones no dan créditos a mujeres o estiman que los varones negros o hispanos son más propensos a delinquir que los varones blancos. «El sesgo impregna nuestro mundo», afirma certeramente Coeckelbergh (2021, 113).

Otra consecuencia indeseable del sesgo en los datos son las burbujas de filtro, inducidas por los algoritmos que nos sugieren enlaces al navegar por Internet tras clasificarnos por el vecino más cercano (aquel que haya dado nuestros mismos clics o haya realizado búsquedas similares, de modo que si éste pinchó en X, se nos ofrecerá X para pinchar). Los filtros son, desde luego, útiles, al personalizar los contenidos que se nos van apareciendo en función de nuestros intereses (del mismo modo que es útil el filtrado bayesiano del correo basura en nuestra bandeja de entrada). Pero conllevan riesgos, como la creación de burbujas en torno al usuario por las que termina viendo únicamente contenido que se ajuste a sus propios valores y selecciones previas, lo que puede ser enormemente perjudicial. Pensemos, por ejemplo, en la información política. Si el buscador de Google o el algoritmo de Facebook nos muestran siempre páginas o noticias con el mismo sesgo político, al tiempo que Youtube nos ofrece vídeos en sintonía, se corre el riesgo de que el usuario quede encapsulado en sus creencias y opiniones, viéndolas reforzadas una y otra vez, sin llegar a cuestionárselas nunca. Esto puede provocar que la sociedad tienda a polarizarse cada vez más. Es, por tanto, conveniente que los usuarios borren su historial periódicamente, o naveguen en oculto, a fin de evitar que los algoritmos los encasillen. Y, por parte de las empresas, es conveniente que los algoritmos incluyan componentes estocásticos, aleatorios, de modo que no siempre recomienden el producto que consumió el vecino más cercano, los usuarios con gustos similares.

Pasemos ahora al sesgo de los algoritmos. Suele decirse que lo que tiene sesgo no son los algoritmos de IA sino los datos, queriendo con ello excusar a los algoritmos (y sus programadores): si un algoritmo produce predicciones sesgadas, es por el sesgo presente en los datos de entrenamiento. Un detalle para nada baladí es que cuando un sistema con IA no funciona como se esperaba, los programadores

con frecuencia echan la culpa a los datos, al sesgo que arrastraban, y, sin embargo, cuando el sistema funciona correctamente, no ponen de relieve la importancia de los datos sino la inteligencia que el sistema tendría por sí mismo.

Pero claro que hay un sesgo algorítmico: el sesgo algorítmico tiene que ver con la selección del tipo de algoritmo, o de las variables y los métodos aplicables, siendo el sesgo por los datos de entrenamiento posterior. Se habla de sesgo en el algoritmo cuando el modelo realiza asunciones erróneas. Por ejemplo, planteamos un modelo de regresión para estudiar la evolución de la temperatura global anual del siglo XX y elegimos como variable predictora el número de piratas en el mundo en lugar de la concentración de gases de efecto invernadero, la radiación solar u otros factores (el caso es que el modelo será exitoso, porque el resultado es una correlación inversa fuerte; pero espuria, al no haber conexión causal entre la temperatura global y el número de piratas en el mundo). En cualquier caso, el sesgo algorítmico es, en cierto sentido, inevitable, porque el algoritmo del sistema de IA está diseñado para aprender algo (ajedrez) pero no otra cosa (damas).

Un ejemplo curioso de sesgo algorítmico lo ofrece ChatGPT, en el cual, a diferencia de Tay, debe de haber una función que controla qué textos puede manejar y cuáles no, esto es, cuáles filtra o censura, dependiendo del control ideológico que ejercen sus diseñadores, lo que introduce cierto sesgo en las respuestas que da.<sup>62</sup> Así, el programa puede hacer un poema admirando al presidente demócrata Joe Biden pero no le está permitido hacerle lo mismo al expresidente republicano Donald Trump. De igual modo, ChatGPT no te cuenta un chiste sobre mujeres pero sí lo hace sobre hombres. (Por su parte, DALL-E advierte a los usuarios de un bloqueo permanente si buscan que produzca imágenes sobre ciertos temas peliagudos como el antinatalismo; y los algoritmos de Youtube o Facebook posicionan siempre al lado de cualquier vídeo o publicación que contenga el término «cambio climático» enlaces a la postura oficial

---

62 OpenAI, según ha revelado un reportaje de investigación de la revista *TIME* dado a conocer en enero de 2023, ha empleado a cientos de trabajadores mal pagados en Kenia para reconocer respuestas tóxicas de ChatGPT y etiquetarlas como inapropiadas, lo que refuerza al programa para eliminarlas.



para pastoreo de la masa.) Como recoge Borji (2023, 27), diversos test han mostrado que ChatGPT se inclina por el punto de vista de la izquierda socialdemócrata, a pesar de que lo mínimo que se esperaría de un chatbot es que fuera equidistante.

En suma, considerando la evidencia de que los datos que se usan en IA son siempre parciales, en tanto que no describen la realidad en su totalidad y, por consiguiente, los sesgos en un sentido u otro están garantizados y son inevitables, ¿es racional confiar tanto en la IA para la toma de decisiones? Hay que mejorar tanto en la selección de los datos, pues aunque se disponga de cantidad pueden ser irrelevantes, como en el diseño del modelo y la realización de las inferencias.

c) Explicabilidad de las cajas negras:

Como vimos en el capítulo 1, la IA simbólica (la de los sistemas expertos) era deductiva, de arriba a abajo, basada en reglas a priori. En cambio, la actual IA guiada por los datos (la de las redes neuronales) es inductiva, de abajo a arriba, delimitando reglas a posteriori.<sup>63</sup> El paso de una IA a la otra ha supuesto un cambio en las técnicas de programación: si antes se perseguía construir una máquina computacional que realizase determinada tarea en función del programa introducido, que se programaba explícitamente para ello, ahora se persigue la misma meta pero sin que la máquina esté específicamente programada para ello. Sólo se programa el algoritmo de aprendizaje automático o profundo que desea usarse y se lo entrena con una cantidad astronómica de datos para que «aprenda» a realizar la tarea en cuestión. El programador ya no codifica las reglas que quiere que emplee el sistema, sino que deja que la IA generalice sus propias reglas a partir de la exposición a una batería de ejemplos.

Pues bien, como asevera Génova (2023), la vieja y buena IA simbólica era transparente. Los sistemas con IA lógica eran cajas blancas, que respondían a un diseño racional, en el sentido de intencionado, con una razón de ser (una finalidad propositiva). Las reglas a priori que se usaban para programarlos reflejaban el conocimiento de los expertos. En el caso de un sistema experto empleado en medicina, las reglas implementadas

---

63 También hay una IA híbrida o mixta, que combina ambas, como pasa en los sistemas de traducción automática, que se apoyan tanto en la gramática como en la estadística.

(«si el nivel de colesterol en sangre del paciente supera los 250 mg/dl, el paciente tiene hipercolesterolemia») se basaban en el conocimiento acumulado por los médicos, cayendo por tanto bajo su dominio y responsabilidad. Ahora bien, esto también determinaba que el sistema nunca fuera más allá del saber establecido (no «aprendiera» nada nuevo).

A diferencia, la nueva IA es en múltiples casos opaca. Sistemas como las redes neuronales constituyen insondables cajas negras, cuya confiabilidad queda en entredicho. Las reglas a posteriori que extraen los sistemas dotados de aprendizaje automático se basan, simplemente, en la inducción a partir de las regularidades que observan en los datos. Es decir, la racionalidad de las redes neuronales no se funda en que hayan sido diseñadas específicamente para cumplir tal tarea sino en que, al tener millones de parámetros ajustables, son capaces de ajustarse prácticamente a cualquier conjunto de datos por complejo que sea, no siendo posible en principio ofrecer otra explicación de las reglas a posteriori que inducen. Al no haber un diseño racional (salvo la bondad del ajuste a los datos), no es posible explicar por qué funcionan las reglas que coligen, pero tampoco por qué fallan cuando lo hacen (Génova 2023).

Entonces, si nos volvemos a poner en el contexto de la IA aplicada a la medicina, cabe preguntarse ante el uso médico de un sistema dotado de aprendizaje profundo lo siguiente, a saber: ¿qué médico, ingeniero o programador está en condiciones de hacerse responsable de unas decisiones que no sabe explicar (a pesar de que las recomendaciones resulten plausibles)? Y esto lo preguntamos no sólo con respecto a las decisiones que comporten errores, sino especialmente con respecto a las decisiones que resulten correctas. Porque aceptar sin explicación que el algoritmo funciona mejor es como pedirle al médico que se arrodille ante el curandero. La magia del curandero puede, desde luego, tener su funcionalidad, pero el médico ha de poder explicar en términos científicos (causas y mecanismos) la razón del éxito de la técnica del curandero. Ha de abrir la caja negra, sea la de la técnica mágica del curandero o sea la del sistema con IA, porque el paciente sólo puede dar su consentimiento informado si ha recibido del médico una explicación lo suficientemente detallada de la prueba a la que va a someterse.

Y esto que planteamos con respecto a los algoritmos que realizan diagnósticos clínicos, también cabe plantearlo con respecto a los

algoritmos que deciden inversiones, controlan vehículos, granjas o plantaciones y, por no seguir, los agentes autónomos (como las futuras armas autónomas letales, que escogerán ellas mismas sus objetivos), que podrían tomar decisiones no alineadas con los objetivos humanos.

La opacidad que tiñe las redes neuronales multicapa plantea, pese a su popularidad e innegables éxitos, serias dudas sobre su funcionamiento con relación a la responsabilidad por sus resultados. La IA no puede convertirse en una industria de cajas negras que comportan una inexplicabilidad fundamental; porque no es aceptable la toma de decisiones mediante ingeniería de caja negra, por más que el resultado responda al ajuste a los datos de entrenamiento, pues el ingeniero ha de poder interpretar y explicar (a sí mismo y a los demás) por qué se obtiene precisamente ese resultado concreto y no otro, a pesar de que conozca el funcionamiento general. En último término, dado que no resulta posible interpretar y explicar cómo se obtienen los *outputs* a partir de los *inputs*, la responsabilidad en caso de mal funcionamiento recae en los ingenieros y programadores que optaron conscientemente por implementar una red neuronal profunda en lugar de otro método de IA más transparente (Coeckelbergh 2021, 103).

No obstante, hay que advertir que parte del problema radica en que los métodos con mayor capacidad predictiva (redes neuronales, máquinas de soporte vectorial, &c.) son, por su no linealidad y gran complejidad computacional, los métodos con menor explicabilidad. Y, recíprocamente, los métodos más explicables (regresión, árboles de decisión, &c.) son, por su linealidad y poca complejidad computacional, los métodos con menos rendimiento predictivo. A diferencia de los árboles de decisión o la regresión (lineal, polinómica, logística), donde es posible saber cómo el algoritmo toma la decisión que toma en función de unas reglas de decisión y unos pesos sobre las variables de entrada (aunque si estamos ante un árbol gigantesco, recorrer el camino desde el nodo inicial hasta el nodo final puede ser fatigoso), el gran número de operaciones internas cruzadas de una red neuronal multicapa hace de raíz muy complicado interpretar cómo afecta cada variable de entrada a la decisión final (pudiendo producirse lo que en teoría del caos se denomina efecto mariposa).<sup>64</sup>

---

64 Jugando con los valores de las variables de entrada, podemos intentar hacernos

En ocasiones, se distingue entre interpretabilidad y explicabilidad (CoboCano & Lloret Iglesias 2023, 58). Mientras que la interpretabilidad remitiría a la posibilidad de entender la causa del resultado producido por un algoritmo, la explicabilidad referiría a la capacidad de ofrecer una explicación comprensible del funcionamiento del algoritmo. Pero, a nuestro juicio, la explicabilidad implica la interpretabilidad, pues uno no puede explicar lo que no sabe interpretar.

d) Implicaciones sociales (*nuevas mentirosas*, desempleo...):

Aparte de las implicaciones sociales relacionadas con la privacidad, los sesgos y la explicabilidad que hemos apuntado más arriba, el auge de la IA comporta otra serie de implicaciones sociales que implican cambios a escala ética y, en especial, moral y política, afectando a los grupos y los Estados.

Nos referimos, por un lado, a las *nuevas mentirosas* o paparruchas («fake news»). Actualmente, hay programas como MidJourney, Face2Face o Lyrebird capaces de imitar verosímilmente la cara, las expresiones faciales y la voz, respectivamente, de una persona para crear imágenes, vídeos y audios donde la persona hace o dice cosas que no ha hecho o dicho, lo que permite la producción y transmisión de información dudosa y noticias falsas.<sup>65</sup> La viralidad de estas nuevas mentirosas ha dejado ya su huella en las elecciones presidenciales de EE.UU. de 2016, el Brexit o el golpe de Estado secesionista catalán de 2017.

Sin embargo, los problemas relacionados con la desinformación, así como con la protección de los derechos de autor, que plantea la generación de textos, imágenes y vídeos mediante IA no son radicalmente nuevos, porque se asemejan a los que en su día planteó otra tecnología: la imprenta, que favoreció el plagio y la difusión de mentiras (por ejemplo, Teodoro de Bry incluyó, en latín, en su obra *América* fragmentos de la *Historia natural y moral de las Indias* del

---

una idea de su relevancia observando cómo afectan a la predicción final de la red neuronal; pero si entre las variables de entrada existen correlaciones ocultas (lo que en estadística se llama multicolinealidad), comprender la importancia de las variables predictoras en el resultado final puede volver a complicarse.

65 Es posible incluso imitar el estilo musical de un compositor. Así, DeepBach, tras ser entrenado con 352 cantatas del célebre músico, logró componer fragmentos que engañaron a especialistas.

padre Acosta sin que constara su nombre, y la imprenta fue clave en la difusión de la propaganda protestante antiespañola, de la leyenda negra, mediante grabados y textos como el de los De Bry).

Por otro lado, la automatización de ciertas tareas rutinarias o repetitivas puede conllevar cambios drásticos en el mercado laboral. Para Kaplan (2017, 128-132), la IA amenaza al personal de almacén, los conductores, las cajeras de supermercado o los teleoperadores; pero no amenaza al personal de servicios cara a cara, los desarrolladores informáticos o los operadores de robótica encargados del mantenimiento y la asistencia de las IAs.

No obstante, un reciente informe no revisado por pares de OpenAI señala que, dado que los avances punteros están en la dirección de los enormes modelos de lenguaje, los trabajos reemplazables mediante IA van a ser, paradójicamente, aquellos que más tienen que ver con estar delante de una pantalla para generar textos, imágenes o código (Eloundou & *al.* 2023). Los trabajos más propensos a ser digitalizados serían aquellos que precisamente se realizan sentado delante de una pantalla de ordenador (traductores, periodistas, guionistas, diseñadores gráficos y otros puestos de la empresa audiovisual), ya que tratan con entidades informáticas que ChatGPT o DALL-E pueden producir, como bien saben los alumnos de medio mundo. Por el contrario, los trabajos agrícolas o los mineros, así como los que tienen que ver con el pensamiento científico o crítico, serían –según el informe antedicho– los de menor exposición. No deja de ser curioso que la IA vaya (supuestamente) a reemplazar a los humanos antes en tareas propias de las artes liberales que de las artes mecánicas.

De otra parte, es ciencia ficción columbrar que las máquinas sustituirán a los jueces y los abogados en el futuro (como lo es, según defendimos en 3.6, que vayan a hacerlo con los médicos y los enfermeros), porque el Derecho no es, pese a la filosofía espontánea de muchos leguleyos, un sistema axiomático consistente, completo y decidible. Por no ser no es siquiera un álgebra, susceptible de ser computada. En todo caso, se automatizarán ciertas tareas relacionadas con el oficio, a pesar de que es famoso el bufete de abogados estadounidense que empleó ChatGPT para documentar un caso, presentando al juez una serie de casos reales que apoyaban su postura y que, en realidad, la IA generativa se había inventado.

En suma, más que una pérdida de puestos de trabajo, se van a producir cambios en su distribución, sin que ello necesariamente implique que el número total permanezca constante.<sup>66</sup>

#### 5.4 Indefinición de los marcos normativos vigentes

El cúmulo de «dilemas éticos» que hemos esbozado ha motivado que múltiples organismos nacionales e internacionales hayan aventado marcos éticos y normativa encaminada a legislar sobre la protección de datos, la equidad y no discriminación de los algoritmos, la seguridad, la transparencia y la rendición de cuentas de los sistemas con IA, la sostenibilidad ecológica... Como sostiene Coeckelbergh (2023, 17), la IA es una tecnología política de cabo a rabo, pues no es políticamente neutral, al afectar a todas las capas de las sociedades políticas.

En Europa, la Comisión Europea organizó en junio de 2018 el Grupo Independiente de Expertos de Alto Nivel sobre Inteligencia Artificial (del que formaba parte Luciano Floridi, según relata en Floridi (2021)), que publicó en 2019 *Directrices éticas para una IA fiable* y *Recomendaciones de políticas e inversiones para una IA fiable*. Basándose en estos documentos, la Comisión Europea preparó en febrero de 2020 el *Libro Blanco sobre la IA*, enfocado a la mitigación de los riesgos inherentes a la IA (Comisión Europea 2020). Y el 21 de abril de 2021 la Comisión Europea presentó un borrador de *Acta sobre la IA para la Unión Europea*, enmendado y acordado el 8 de diciembre de 2023 y en vías de ratificación (Comisión Europea 2021). Las líneas maestras de esta futura ley europea sobre la IA, pionera en el mundo y que se espera que esté lista en 2024 y entre en vigor en 2025 ó 2026, son, en consonancia con

---

66 Algún lector puede echar en falta que no discutamos el dilema ético del tranvía en relación con los coches autónomos, que ha hecho correr ríos de tinta y consiste en la siguiente situación fatal: un coche autónomo debe optar inevitablemente entre atropellar a un peatón y salvar a su conductor o esquivar al peatón y matar al conductor al salirse de la carretera por un terraplén. A nuestro entender, se trata de un pseudodilema, tan improbable que se dé exactamente en la realidad como que un asno muera por tener dos cubos de heno a la misma distancia (el asno de Buridán).

lo que venimos discutiendo: proteger la privacidad y la gobernanza de los datos (prohibiendo la identificación biométrica remota y considerando los sistemas de recomendación de contenidos como sistemas de riesgo); regular las IAs generativas (defendiendo los derechos de autor de las fuentes –razón por la que Sam Altman, CEO de OpenAI, barajó la posibilidad de que ChatGPT o DALL-E no dieran servicio en Europa– y obligando a etiquetar los contenidos así generados); apostar por la seguridad, el control de sesgos y la transparencia (aunque ello vulnere el secreto comercial de las empresas); y buscar el bienestar social y ambiental.

Es realmente interesante atender a los fundamentos filosóficos de los documentos que ha ido produciendo la Unión Europea, pues las ideas filosóficas que actúan detrás son, en ocasiones, muy oscuras y confusas. Conviene recordar, para que se vea que lo expuesto en 5.2 no es una discusión bizantina, que, en 2017, el Parlamento Europeo trasladó (sin éxito) a la Comisión Europea una resolución en la que recomendaba aceptar que ciertos robots inteligentes eran, a efectos jurídicos, «personalidades electrónicas», responsables de sus actos y de reparar los daños que puedan causar, con derechos y deberes y la posibilidad de firmar contratos (una especie de Proyecto Gran Robot a la hechura del Proyecto Gran Simio, donde los «brutos» del presente en marcha ya no serían los animales sino los robots). Este tipo de deslices es lo que hace tan importante definir con exactitud qué es y qué no es la IA, porque la definición que finalmente sea aceptada delimitará el alcance de las leyes de la Unión Europea. Al respecto, el Acta sobre la IA en trámite establece que un sistema de IA significa un sistema basado en una máquina diseñada para operar con diferentes niveles de autonomía y que puede, para objetivos explícitos o implícitos, generar resultados tales como contenidos, predicciones, recomendaciones o decisiones, que influyen en entornos físicos y virtuales (Comisión Europea 2021, artículo 3).

Según la Comunicación «IA para Europa» de la Comisión Europea al Parlamento Europeo y al Consejo Europeo del 25 de abril de 2018, la Unión Europea persigue ser competitiva en el campo de la IA, atrayendo inversiones privadas (en las que se sitúa a la zaga), pues –como recogen las páginas 4 y 5 de la Comunicación– EE.UU. está invirtiendo mucho dinero y China también lo está haciendo en ingentes

cantidades, buscando alcanzar el liderazgo mundial en la década de 2030 con su «Plan de desarrollo de la IA de próxima generación», así como Japón y Canadá cuentan con estrategias al respecto. Pero la Unión Europea lo quiere hacer de modo que «las nuevas tecnologías estén basadas en valores» y no generen «problemas de carácter ético y jurídico». En 2019, el Grupo de Expertos de Alto Nivel sobre IA presentó unas *Directrices éticas para una IA fiable* (Grupo de Expertos de Alto Nivel sobre IA de la Unión Europea [GEANIA] 2019a), como alternativa a los marcos éticos propuestos por otros organismos (y que más abajo analizamos). Este grupo entiende como IA fiable una IA lícita, ética y robusta (técnica y socialmente), es decir, una IA que respete las leyes y los reglamentos, así como los principios y los valores, y que no provoque daños (GEANIA 2019a, 2).

En lo que toca al aspecto ético, el grupo entiende por ética de la IA una parte de la ética aplicada, que a su vez sería una parte de la ética académica en cuanto subdisciplina de la filosofía y dividida en metaética, ética normativa, ética descriptiva y ética aplicada (GEANIA 2019a, 11 y 49). Por ética comprenden el estudio académico de la moral, esto es, de las pautas de comportamiento, aferrándose a la visión intelectualista de la ética que criticamos al inicio del capítulo:

En los debates éticos se utilizan con frecuencia los términos «moral» y «ético». El primero se refiere a lo concreto, las pautas de comportamiento, las costumbres y convenciones que se pueden encontrar en determinadas culturas, grupos o personas en un momento específico. El término «ético», por su parte, hace referencia a una evaluación de esas acciones y comportamientos concretos desde una perspectiva sistemática y académica (GEANIA 2019a, 49).

Obsérvese, además, cómo equiparan todas las normas, dando igual que provengan de «culturas, grupos o personas», es decir, sin distinguir entre normas políticas (sociedades), morales (grupos) y éticas (individuos), lo que acarreará consecuencias para la tecnoética que esbozan. Esta visión de la ética más dialógica que operatoria es, por demás, la que está detrás del constitucionalismo digital de inspiración habermasiana hegemónico en las instituciones de la Unión Europea.



A su entender, una IA ética ha de estar basada en los siguientes cuatro principios éticos, que parecen entresacados de códigos bioéticos (salvo el último): respeto de la autonomía humana, prevención del daño, equidad y explicabilidad (GEANIA 2019a, 14). Desde nuestras coordenadas, conforme a la tecnoética materialista, sólo los dos primeros principios serían éticos, al tener que ver con la firmeza del individuo. No obstante, cuando el daño al que hace referencia el segundo principio repercute en grupos («los daños pueden ser individuales o colectivos»; GEANIA 2019a, 15), estaríamos ante un principio moral. En cambio, los dos últimos principios serían más bien principios políticos, al tener que ver con la justicia social y con los perjuicios sociales que puede ocasionar una IA de caja negra, afectando a las sociedades políticas.

Es notorio que los «expertos independientes de alto nivel» no distinguen, empero, entre ética, moral y política. Una indistinción que produce puntos de fricción en sus directrices. Así, los propios expertos se dan cuenta de que algo no marcha bien cuando reconocen «tensiones entre los diferentes principios» (GEANIA 2019a, 16). Por ejemplo, la prevención del daño mediante sistemas de IA para la actuación policial predictiva (como la identificación y el seguimiento de personas mediante IA, o la evaluación de los ciudadanos por puntos como en China) puede entrar en contradicción con la autonomía de la persona (con su derecho a la privacidad). No columbran, en suma, que la ética puede chocar y choca con la moral y la política, y viceversa (tal y como expusimos). La unidad de los cuatro principios postulados es falaz, porque el sistema contiene en su seno discontinuidades radicales, conformando lo que en matemáticas se llamaría un sistema axiomático inconsistente, puesto que los principios se contradicen entre sí.

Ahora bien, ¿cuál es el fundamento de esos principios? Los autores indican que los cuatro principios están «arraigados en los derechos fundamentales», consagrados en los tratados de la Unión Europea, como la Carta de Derechos Fundamentales, y en la legislación internacional de Derechos Humanos (libertad del individuo, respeto por la democracia, &c.), y cuya base común es «el respeto de la dignidad humana», es decir, del valor que se presupone intrínseco e igual para cada ser humano con independencia de su raza, color,

ascendencia, género, edad, idioma, religión, opiniones políticas, origen nacional, étnico o social, condición económica o social de nacimiento, &c. (GEANIA 2019a, 12 y 49). Pero si al ser humano le borramos su raza, color, género, idioma, nacionalidad, &c., pues nos quedamos sin ser humano y más bien obtenemos un australopiteco o un ángel, que carecen de nacionalidad o sexo, respectivamente. La dignidad humana es una noción cuyo poso teológico muchas veces se olvida, pues tiene que ver con que Dios se hizo hombre, se encarnó en Jesucristo, lo que defiende la dignidad del hombre cristiano frente a los animales y los ángeles pero también frente a los judíos y los musulmanes. Resumiendo, tenemos que la IA ética descansa en unos principios cuyo fundamento (metafísico) son, en el fondo, los Derechos Humanos y, si les apuran, la dignidad humana. Estamos ante una concepción de los Derechos Humanos como una suerte de derecho natural emanado del Hombre, de la Humanidad, como otrora emanaba de Dios. Pero todo derecho es siempre derecho positivo, no existiendo un único derecho sino múltiples derechos, pues cada Estado tiene el suyo, entrando en contradicción unos con otros. Y el fundamento del derecho no es otro que la fuerza relativa de cada Estado para imponer el suyo sobre los demás (Martín Jiménez 2021). Los Derechos Humanos de 1948 son el derecho de EE.UU. y sus aliados tras la II Guerra Mundial, pues nunca los votó la U.R.S.S. o los ha ratificado China, y los países árabes tienen su propia declaración de Derechos Humanos en el Islam (la Declaración de El Cairo, de 1990, cuyo fundamento es la Sharía).

Los principios propuestos por el grupo, concebidos con resonancias kantianas como «imperativos éticos» (GEANIA 2019a, 14), estarían orientados a lograr los omnipresentes objetivos de desarrollo sostenible de la Agenda 2030 (GEANIA 2019a, 11), y así conseguir un futuro con mayor bienestar individual y colectivo, una sociedad más justa, en la que reinen la Democracia y el Estado de Derecho (como si todo Estado no fuera Estado de Derecho *per se*).

Desde un punto de vista más prosaico, el Grupo de Expertos ha producido una Lista de Evaluación de la IA fiable (conocida como ALTAI) con 7 ítems que funcionan a modo de requerimientos éticos que las empresas han de cumplir a la hora de desarrollar sus sistemas de IA (GEANIA 2020):

1. Respeto de la acción y supervisión humanas.
2. Solidez técnica y seguridad.
3. Protección de la privacidad y buen gobierno de los datos.
4. Transparencia.
5. No discriminación y equidad.
6. Bienestar social y ambiental.
7. Rendición de cuentas.

Pero, de nuevo, estas recomendaciones, como los principios que las informan, resultan demasiado abstractas, porque: ¿cómo asegurar la supervisión humana del sistema, no ya en el caso de que el ser humano intervenga en todo el ciclo de decisión del sistema (*human in the loop*), sino en los casos de que sólo intervenga en su monitorización (*human on the loop*) o, lo que es más difícil, de que no sea necesaria ninguna acción suya (*human out the loop*)? ¿Es técnicamente posible garantizar la seguridad de los sistemas con IA? ¿Cómo proteger de facto la privacidad de los ciudadanos? ¿Cómo controlar los sesgos en los datos y en los algoritmos? ¿Cómo asegurar su explicabilidad sin quebrantar el secreto empresarial? ¿Cómo establecer la cadena de responsabilidad en caso de fallo? ¿Cómo evitar que el auge de la IA deje sin empleo a muchos trabajadores o perturbe las campañas electorales? ¿Cómo avanzar de un modo ecológico y sostenible en el desarrollo de esta tecnología que paralelamente consume tantos recursos energéticos y minerales?

Si dejamos la Unión Europea y nos fijamos en otros organismos, nos encontramos con las propuestas de la OCDE (*Recomendación sobre IA*, 2019) y de la UNESCO (*Recomendación sobre la Ética de la IA*, 2021).<sup>67</sup> En ellas, hallamos la misma confusión entre principios éticos, morales y políticos, así como la apelación a la dignidad humana y los Derechos Humanos como fundamento último de los mismos. Los cinco principios sobre IA de la Organización para la Cooperación y el Desarrollo Económico (OCDE) fueron el primer marco intergubernamental adoptado. El segundo principio apela a los valores humanos, esto es, a los Derechos Humanos y los valores democráticos (como puede leerse en la página 7 del documento). La

---

67 Consúltense <https://oecd.ai/en/ai-principles> y <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) formula su Recomendación como un instrumento «centrado en la dignidad humana y los derechos humanos, así como en la igualdad de género, la justicia social y económica y el desarrollo, el bienestar físico y mental, la diversidad, la interconexión, la inclusión y la protección del medio ambiente y de los ecosistemas» (esta rapsodia de buenos deseos puede encontrarse en la primera página). Y, más adelante, en su página 6, expresa como primer objetivo «proporcionar un marco universal de valores, principios y acciones para orientar a los Estados»; pero que exista ese marco universal es precisamente lo que está en discusión, porque toda tabla de valores de un grupo o una sociedad tiene enfrente otra tabla de contravalores seguida por otros grupos o sociedades (como vimos, las normas éticas son universales, pero las normas morales y políticas no). Así, en la página 21, la UNESCO apuesta por la elaboración de políticas con perspectiva de género, destinadas a que las niñas y las mujeres no queden excluidas de la economía digital impulsada por la IA; pero cabe dudar de que todos los países musulmanes velen por ello.

Al otro lado del Atlántico, tenemos los *23 Principios de la IA de Asilomar* (2017), planteados en la reunión concertada en ese lugar de California por el Instituto para el Futuro de la Vida (cuyo presidente es el cosmólogo Max Tegmark y cuyo cofundador es el programador Jaan Tallinn). Estos principios cuentan con el beneplácito de investigadores y empresarios del campo como Stephen Hawking, Elon Musk, Demis Hassabis, Yann LeCun, Yoshua Bengio, Ray Kurzweil o Stuart Russell, y desarrollan las tres leyes de la robótica que planteara Isaac Asimov (en el Anexo III ofrecemos su traducción al español, ya que en la página del Instituto para el Futuro de la Vida sólo aparecen en inglés, chino, alemán, japonés, coreano y ruso). El undécimo principio vuelve a hablar de «valores humanos», remitiendo a la dignidad humana y los Derechos Humanos. De otra parte, abundando en la confusión, la *Declaración de Montreal para un Desarrollo Responsable de la IA* (2018) hace referencia en su primer principio –principio de bienestar– al cultivo del bienestar de todos los «seres conscientes», un cajón de sastre donde entrarían los hombres y, para algunos, los animales e, incluso, los robots inteligentes. Otra muestra de la indefinición reinante es el sexto principio o principio de equidad, que establece que «el desarrollo y el

uso de los SIA [Sistemas de Inteligencia Artificial] deben contribuir a crear una sociedad justa y equitativa» (algo parecido exhortan los mandamientos 14, 15 y 17 de Asilomar); pero a qué sociedad se refieren, ¿a Canadá, EE.UU., Europa, India o China? Porque el desarrollo y el uso beneficioso de los SIA en el seno de un Estado bien puede ir en contra de la eutaxia de otros Estados, si, pongamos por caso, se hace más competitivo económicamente.

En resumen, el análisis del casi centenar de guías éticas de la IA disponibles muestra que a menudo se repiten principios como transparencia, justicia y equidad, no maleficencia, responsabilidad y privacidad (Jobin & *al.* 2019), aunque cada organismo los entienda a su manera y de forma no siempre compatible porque –como explicamos– no existe una única tecnoética sino una pluralidad polémica de ellas. Además, según el mapa comparativo entre las distintas listas de principios propuestos en la ética de la IA por las diferentes instituciones, elaborado por el Centro Berkman Klein de la Universidad de Harvard, la mayoría adopta explícitamente o hace referencia a la Declaración de Derechos Humanos (22 de las 35 declaraciones de principios).<sup>68</sup>

Concatenar el adjetivo «ética» con IA funciona muchas veces como concatenar el adjetivo «verde» con tecnología, esto es, como un lavado de cara. Cuando se leen las guías para una IA ética de algunas empresas se comprueba que nunca falta toda la cháchara eticista («una IA fiable y confiable, centrada en el ser humano, inclusiva, que respete la diversidad...») y que en ocasiones hasta ciertas partes han sido redactadas por ChatGPT (como sucede con la hoja de ruta de la EASA para la IA en aviación; EASA 2023, 7). Una mejora de la imagen más que una llamada en serio a domesticar la IA a la manera que se pusieron límites de velocidad a los coches o se reglamentó el tráfico para hacer la conducción más segura y ordenada.

## 5.5 Geopolítica de la IA

La mayoría de los marcos normativos vigentes se inspiran, según sus exposiciones de motivos, en la llamada «ética de la IA», cuyas reglas

---

68 [https://wilkins.law.harvard.edu/misc/PrincipledAI\\_FinalGraphic.jpg](https://wilkins.law.harvard.edu/misc/PrincipledAI_FinalGraphic.jpg)

consideran universales al apelar en último término a los Derechos Humanos y la dignidad del hombre. Pero, como explicamos, la tecnoética es también tecnomoral y tecnopolítica. En realidad, las diferentes normativas (de la Unión Europea, de la OCDE, de la UNESCO...) están codeterminadas por las circunstancias económicas, sociales e históricas de cada sociedad política. Con otras palabras: no se puede hablar de «ética de la IA» sin hablar de (geo)política de la IA o, por mejor decir, sin hablar de la dialéctica de Estados e Imperios con relación a esta tecnología. Como dijo Vladimir Putin en septiembre de 2017 de manera hiperbólica, «el país que domine la IA dominará el mundo».

De hecho, a finales de octubre de 2023, el G7, conformado por Estados Unidos, Reino Unido, Japón, Canadá, Alemania, Francia e Italia, anunció que ya tenía acuerdo sobre la IA, habiendo llegado a un compromiso –la tercera vía propuesta por Japón– entre las posiciones antagónicas al respecto de EE.UU. («no poner trabas a su desarrollo») y Europa («regularla de manera estricta»). Pero este acuerdo en regular la IA, aunque sea de manera laxa, es voluntario. Porque ninguno de los países del G7 tiene en verdad la obligación de transformarlo en ley. EE.UU. prima el interés de sus grandes compañías tecnológicas en aras de su eutaxia, mientras que la Unión Europea es, por prudencia política, partidaria de regular la injerencia de la IA foránea (y esta puede ser la funcionalidad real del eticismo de que hace gala Europa).

De la importancia de la tecnopolítica habla que, para algunos Estados, los robots inteligentes constituyen sujetos políticos. Por un lado, en octubre de 2017, Arabia Saudí otorgó la ciudadanía a la ginoide Sophia, convirtiéndose en el primer robot con nacionalidad. Por otro lado, en abril de 2018, otro androide con rasgos femeninos, bautizado como Michihito Matsuda, fue presentado como candidato a las elecciones municipales en un distrito japonés y, con su promesa de acabar con la corrupción y las malas decisiones políticas mediante un algoritmo de *machine learning*, quedó tercero en número de votos.

La tecnoética materialista sustenta que los sistemas con IA no son sujetos políticos, porque ni siquiera son sujetos éticos, por las razones que expusimos. La otorgación de la ciudadanía saudí a Sophia no fue sino un montaje mediático auspiciado por la empresa fabricante y, por su parte, detrás de Michichito se encontraba un alto cargo de un proveedor de servicios móviles y un expleado de Google Japón.

Con estos dos ejemplos quiero poner de relieve que detrás de la IA están grupos respaldados por compañías empresariales y/o Estados. Por decirlo con Kate Crawford (2021, 211):

Los sistemas de Inteligencia Artificial están diseñados para ver el mundo e intervenir en él de maneras que beneficien principalmente a los Estados, las instituciones y las corporaciones a los que sirven.

Y en este punto nos encontramos, como primeros de la lista, a los gigantes tecnológicos estadounidenses (Google –Alphabet–, Amazon, Facebook –Meta–, Apple y Microsoft, con frecuencia mencionados mediante el acrónimo GAFAM) y chinos (Baidu, Alibaba, Tencent, Xiaomi, o sea, BATX). Sin perjuicio de las batallas entre las propias empresas tecnológicas –como Microsoft (OpenAI) y Google (DeepMind), que pugnan por controlar la IA, como pugnan GPT-4 y Gemini–, no hay que ser ingenuos y descartar que detrás de las acciones de estos gigantes estén operando muchas veces intereses políticos y militares. Y no sólo en China, donde el gobierno controla TikTok en la sombra y emplea la tecnología de reconocimiento facial y vigilancia inteligente que le suministra Huawei. Como ya avanzamos, el gobierno de EE.UU. inspeccionaba (¿e inspecciona?), por medio de programas como PRISM o MUSCULAR, los datos y metadatos de los usuarios de Google, Microsoft, Facebook o Youtube. Tras el 11S, el gobierno estadounidense fortaleció sus vínculos con estas compañías a fin de aumentar la vigilancia antiterrorista. La Ley Patriota, aprobada en 2001 después de los atentados, dio permiso a las agencias de seguridad estadounidenses para acceder a los servidores de las empresas tecnológicas con el propósito de combatir el terrorismo. Estamos ante lo que algunos –como la socióloga Shoshana Zuboff– han denominado el capitalismo de la vigilancia (y el comunismo de la vigilancia).

No en vano, los primeros en financiar la IA saliente de la conferencia de Dartmouth de 1956 fueron los militares, a través de ARPA (la Agencia de Proyectos de Investigación Avanzados, fundada en 1958; desde 1972, DARPA, Agencia de Proyectos de Investigación Avanzados en Defensa) (Kaplan 2017, 21-22). Al Pentágono le interesaba invertir en el diseño y la fabricación de máquinas capaces de realizar automáticamente tareas de inteligencia, como transcribir

conversaciones telefónicas pinchadas en ruso y traducirlas al inglés, o como reconocer lanzamisiles, tanques o barcos en fotografías aéreas. Con este objetivo, el Grupo de IA del MIT (al que pertenecían John McCarthy y Marvin Minsky, y que se transformaría en 1968 en el Laboratorio de IA del MIT) recibió importantes sumas de dólares a partir de 1963. ARPA fue también responsable de los fondos que condujeron a ARPANET, la red de ordenadores precedente de Internet. Y conviene recordar que los militares también estuvieron detrás del desarrollo de las primeras computadoras, como el ENIAC, cuyo objeto era agilizar el cálculo de las trayectorias de disparo.

Aún más: aunque tienda a verse la IA como una tecnología meramente formal o abstracta, de igual modo que no pueden soslayarse los planes políticos –respecto a las personas– que operan detrás, tampoco pueden perderse de vista los programas políticos –respecto a las cosas– que están en marcha. Nos referimos a la dependencia que la IA muestra de infraestructuras físicas que están remodelando el mundo, empezando por la explotación de recursos materiales como la energía o el litio que precisan las baterías de los dispositivos. El enfrentamiento entre Estados Unidos y China pasa por hacerse con componentes imprescindibles para la industria tecnológica puntera como los semiconductores y las tierras raras (de las que China sigue siendo el mayor importador mundial, a pesar de disponer de las mayores reservas de dichos minerales; Degli-Esposti 2023, 79).

Frente al mito de la IA como «tecnología limpia», la IA es un sistema extractivo que depende de la explotación depredadora de recursos energéticos y minerales. Crawford (2021, cap. 1) encarna la IA, vinculándola al *mundus adspectabilis*, al mostrar su dependencia de cosas tan prosaicas como la minería natural y la minería de datos. La IA es humana, demasiado humana. Dicho quede, a pesar de que no sea ésta la manera en que la historia de la IA es habitualmente contada.

Pese a que la IA frecuentemente se presente como una palanca esencial en la lucha contra el cambio climático (así lo hace, por ejemplo, el *Libro Blanco sobre la IA* de la Unión Europea, que expone, en su primer párrafo, que la IA contribuirá a la mitigación del cambio climático; Comisión Europea 2020, 1), argumentando que los métodos de IA mejorarán la eficiencia energética o la gestión logística, la IA comporta un impacto climático y una huella ecológica no desdeñable. Entrenar un enorme



modelo de lenguaje, por ejemplo, conlleva un consumo de energía y unas emisiones de gases de efecto invernadero desproporcionadas.<sup>69</sup>

Y, entonces, claro, con este trasfondo político de choque de placas continentales (EE.UU., China, Rusia...) que dibuja el materialismo filosófico, cabe sospechar, en primer lugar, de toda defensa sin fisuras de la IA y los procesos de automatización y robotización que estarían iniciando la llamada «cuarta revolución industrial», tras las provocadas por la irrupción de la máquina de vapor, la electricidad y la electrónica. Precisamente, ha sido el economista Klaus Schwab, fundador y presidente del Foro Económico Mundial (Foro de Davos), quien ha popularizado la expresión, mediante un libro publicado en 2016, para referirse a las innovaciones que en el capitalismo está impulsando la IA, la robótica, el Internet de las cosas y la realidad virtual.

Contra la armonía que las élites globalistas presuponen en el desarrollo del capitalismo, hay que subrayar que el capitalismo internacional no es la suma de los capitalismos nacionales, porque los capitalismos nacionales conforman una biocenosis donde muchas veces los Estados han de doblegar, para mantener su eutaxia, los intereses de multinacionales o gigantes tecnológicos alineados con otro Estado. Tanto el liberalismo como el socialismo realizan una apelación mítica del capitalismo (e, inversamente, del comunismo), aunque el primero lo haga con tintes meliorativos y el segundo peyorativos. Ambas ideologías sustentan el capitalismo, adjudicándole una identidad y una unidad propias. Pero no existe el sistema capitalista, un único sistema capitalista, sino múltiples sistemas capitalistas en *symploké*, formando una pluralidad discontinua. Por la dialéctica de Estados, los diferentes sistemas capitalistas no están acompasados, en sintonía, sino en lucha unos con otros.<sup>70</sup>

---

69 Sobre el cambio climático, véase Madrid Casado (2020).

70 La misma ingenuidad de fondo se detecta en los presupuestos anarcoliberales de que hacen gala muchos analistas del bitcoin y las criptomonedas, que auguran la caída de los Estados (D. María 2024). Más que una moneda que amenace a las monedas estatales son, a día de hoy, activos digitales para invertir, como el arte o los paraísos fiscales. La presunta libertad que ofrece bitcoin a sus poseedores desaparecerá en cuanto amenace la eutaxia de los Estados, que de hecho ya están obligando a declarar su posesión y pueden encarcelar a sus poseedores. Por otra parte, su marcada volatilidad no es una virtud sino un efecto de sus crisis recurrentes; porque es caro y costoso realizar transacciones y, en determinados periodos, se gasta más en energía eléctrica que lo que se gana en minar un bitcoin.

Esta nematología de que estamos a las puertas de un mundo nuevo, de que la historia se está acelerando, cada vez ejerce más influencia en la educación. El último principio ético marcado por la Unión Europea, el principio de explicabilidad, tiene que ver con «no permitir que nadie quede rezagado»... formando «competencia digital, espíritu empresarial y creatividad» en educación (Comisión Europea 2018, 14). Un consejo repetido por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial en sus *Recomendaciones de políticas e inversiones para una IA fiable* de 2019, que aconsejan rediseñar los sistemas educativos desde preescolar a la educación superior. Es así que, con los Fondos Europeos, se está obligando a los profesores a certificar su nivel de Competencia Digital Docente (DigCompEdu), pues parece que si el profesor de matemáticas no enseña a sumar fracciones mediante un recurso digital interactivo que empodere al alumnado sino mediante tiza, pizarra, lápiz y papel, es un mal docente (a pesar de que cada vez más estudios indiquen que lo analógico funciona mejor que lo digital en el aprendizaje; Martí 2022). En lo que afecta a los alumnos, esto se concreta en un hacer sin saber, donde las nuevas tecnologías terminan transformándose para ellos en auténticas cajas negras (alumnos que realizan las tareas en casa recurriendo a ChatGPT o copian en los exámenes usando herramientas con IA). Pero la chatarra digital es, como salía no hace mucho en la prensa, para el vulgo y su prole, porque los gurús digitales de Silicon Valley mandan a sus hijos a colegios donde la tecnología no tiene cabida, donde no se usan móviles, tabletas, ordenadores ni pantallas.<sup>71</sup>

Además, en este contexto de auge de las tecnologías digitales, donde se pretende hacer todo en línea, desde la enseñanza a la atención médica, el trabajo, encontrar pareja o el sexo («cibersexo»), se potencia al consumidor satisfecho con su felicidad canalla, generando individuos digitalmente competentes pero flotantes en el Metaverso.<sup>72</sup> Marino Pérez

---

71 En el cuento *La sensación de poder* (1958), Isaac Asimov imagina un mundo (no tan lejano) en que los hombres ya no saben calcular porque lo hacen los robots, pero un ingeniero ha logrado desmontar uno y, por ingeniería inversa, aprende a multiplicar, lo que se contempla como un descubrimiento revolucionario.

72 El Metaverso, que no avanza tan rápido como esperaba Mark Zuckerberg, se basa en la tecnología de visores inmersivos que suministran al consumidor apariencias falaces. Pero la posibilidad remota de que el Metaverso sustituya al Universo olvida que la identificación de, por ejemplo, una pelota en el Metaverso depende de la previa identificación de una pelota como tal en el *mundus adspectabilis*. Por esta

Álvarez (2023, Parte III) recoge la noción de individuo flotante forjada por Gustavo Bueno (1982), poniendo de relieve la pertinencia que tiene en nuestros días. Bueno (1982, 23) acuñó la noción para referirse a esos individuos cuyos fines desfallecen en relación a los planes y programas colectivos. Y Pérez Álvarez subraya que la hegemonía de lo digital, tanto en el ámbito laboral (el teletrabajo) como en el ocio (las redes sociales), exacerba el número de individuos flotantes, la muchedumbre solitaria (el «solos juntos»). Personas que mayoritariamente sólo interactúan con otras personas a través de experiencias virtuales, controladas por algoritmos de IA que cada vez conocen mejor sus gustos o aversiones, brindándoles sólo aquel contenido que no pincha la burbuja que les rodea y aísla del resto.

Volviendo al principio: la IA no es neutra ética, moral ni políticamente; pero no por ello es inmediatamente una bendición o una amenaza. Será una cosa u otra en función de la tecnología concreta, los individuos, los grupos o las sociedades políticas que tomemos como parámetro. En conclusión, no es, como ingenuamente cree Yuval Noah Harari (2019a, 361; 2019b, 91), que si las decisiones de las máquinas son menos estúpidas que las nuestras (dado que «algoritmos no conscientes pero inteligentísimos pronto podrían conocernos mejor que nosotros mismos»; Noah Harari 2019a, 431), les cederemos la decisión, sino que se la estaremos cediendo *de facto* a los grupos de no tan estúpidos que están detrás de los sistemas con IA. No es, insisto, que las máquinas vayan a controlar a los hombres, sino que ciertos hombres van a controlar a otros hombres a través de las máquinas.

---

razón, Ongay (2024) prefiere conceptualizar la realidad virtual mejor como realidad aumentada (otra hiperrealidad).



## Conclusión

Es momento de poner el punto final o, cuando menos, dada la velocidad a la que progresa la IA, tres puntos suspensivos...

Comenzamos el opúsculo analizando «el mito de la inteligencia artificial», una narrativa escatológica, con raíces en la ciencia ficción, alimentada por los transhumanistas y los futurólogos de la singularidad tecnológica (como Raymond Kurzweil o Nick Bostrom), en la que la IA en ciernes, o bien resolverá todos los problemas, o bien nos aniquilará. Frente al tsunami de noticias y opiniones que diariamente se vierten sobre la IA, ensayamos una filosofía sistemática de la IA desde las coordenadas del materialismo filosófico, articulada en tres partes: gnoseología, ontología y ética de la IA.

En la primera parte tratamos el conjunto de cuestiones que plantean los saberes sobre la IA, proyectando el cono de luz que sobre ellos arroja la teoría del cierre categorial. En este tramo gnoseológico, nos preguntamos si la IA es una ciencia, una técnica o una tecnología. Tras reconstruir brevemente la historia del campo (de la Conferencia de Dartmouth de 1956, donde cristaliza el rótulo «inteligencia artificial» a manos de John McCarthy, pasando por el auge de la IA simbólica, a la explosión actual del aprendizaje automático y el aprendizaje profundo), observamos que algunos autores definen la IA como la ciencia que estudia sistemáticamente el fenómeno de la inteligencia y otros –como Stuart Russell y Peter Norvig– incluso hablan de que se trata de una

ciencia cuyo campo es universal. Desde la teoría del cierre, el campo de la IA no constituye una ciencia sino más bien una tecnología (en rigor, una pluralidad de tecnologías en las que se concitan la lógica, la estadística, &c.), puesto que la trabazón de las operaciones del campo no es interna —estableciendo relaciones entre términos al margen de los sujetos (situación *alfa*)— sino externa (situación *beta*), en el sentido de que las operaciones se traban entre sí al orientarse todas ellas hacia un mismo fin, que suele caracterizarse (*emic*) por construir una máquina capaz de pasar el test de Turing, y nosotros precisamos (*etic*) como diseñar y construir artefactos que realicen determinadas tareas con mayor eficacia y eficiencia que los seres humanos sin su constante supervisión.

A continuación, en la segunda parte, abordamos las cuestiones que plantean, no ya los saberes sobre la IA, sino las realidades de la IA (aunque saberes y realidades son disociables pero inseparables). Dentro del tramo ontológico, nos preguntamos de qué hablamos cuando hablamos de inteligencia y de artificial. Tras describir la filosofía espontánea de muchos especialistas del campo, que funcionan con una visión algorítmica de la inteligencia (sea en la forma de algoritmos lógicos o estadísticos), según la cual la inteligencia humana puede ser formalizada y reproducida fielmente por una máquina, explicamos las diferentes modulaciones de la IA: IA fuerte/débil; IA general/restringida.

Y reconstruimos, desde el materialismo filosófico, los tres argumentos clásicos que se han ofrecido contra las pretensiones de fraguar una IA fuerte y general. Nos referimos al argumento matemático de Lucas-Penrose, el argumento de la habitación china de Searle y el argumento de la informalidad del razonamiento humano (expuesto por Hubert Dreyfus y refinado por Erik Larson). Ahondando en este último argumento, pusimos el foco en que la vieja y buena IA lógica se basaba en la deducción y la actual IA guiada por los datos se basa en la inducción. Pero ni una ni otra logran capturar la abducción, es decir, ese procedimiento contextual por el que los seres humanos conjeturamos explicaciones novedosas o inferimos la causa de un hecho sorprendente. Los algoritmos del *machine learning* detectan patrones en forma de correlaciones a partir de la minería de enormes *datasets*; pero correlación no es causalidad, porque el fundamento de la causalidad son las conexiones, es decir, los mecanismos que actúan debajo de las regularidades estadísticas.

Además, nos detuvimos en el análisis de ChatGPT, una suerte de loro estocástico, que responde a nuestras preguntas aplicando un enorme modelo estadístico de lenguaje, pero que no comprende un ápice de lo que dice. Como vimos, no es tanto que repita textos o devuelva simples yuxtaposiciones, como si fuera exactamente un loro o un mono de repetición, sino que elabora un texto ecléctico, mezclando coherentemente retazos relacionados con el tema de aquí y allá. Pero ChatGPT realiza una manipulación sintáctica del lenguaje, sin asociar ningún tipo de significado semántico, porque las palabras no las vincula a las cosas sino a otras palabras en función de la frecuencia de su aparición concatenada en los textos con que ha sido entrenado.

Sin restar fuerza a los tres argumentos clásicos, señalamos que todos ellos se movían a una escala epistemológica (sujeto/objeto), siendo preciso desbordarla y asumir una escala gnoseológica (materia/forma). La cuestión es que la IA no es, en puridad, inteligencia ni artificial (por contraposición a la humana, que se supone natural). La IA es, por así decir, un baciylmo, una idea desnucada.

Con respecto a la idea de inteligencia, los profesionales del campo de la IA manejan una idea puramente formal, al margen de los materiales. A la manera de los escolásticos, descarnan la inteligencia, reduciéndola a operaciones mentales (formales), no corporales (materiales). Pero no puede haber inteligencia general sin cuerpo. El punto de articulación del sentido común y de la causalidad es la corporalidad: el sujeto corpóreo operatorio. Tanto frente a los que confinan la inteligencia al cerebro humano (los cerebrocentristas), como frente a los que depositan la inteligencia en los ordenadores con independencia de los programadores e ingenieros, el materialismo filosófico sustenta –rectificando dialécticamente ambos extremos– que la racionalidad subjetual y la racionalidad objetual se dan siempre conjugadas, entretrejidas. El mundo está tallado a escala de las operaciones humanas, de manera que los objetos estrasomáticos son objetuales, están ahí, y a la vez son subjetuales, pues por su escala refieren a los sujetos operatorios. Y no es sólo que la inteligencia humana sea abductiva, situacional y operacional, sino que está extendida, externalizada en ingenios objetivos que trascienden positivamente al individuo: las instituciones. Entonces, resulta que la inteligencia humana, que suele calificarse de natural (como si estuviera en los genes), es ella misma artificial, dependiente

del contexto social, cultural e histórico, habiendo tanta «inteligencia artificial» en una mesa, en un teorema matemático, en una sonata... como en ChatGPT.

Por último, en la tercera parte, dedicada a la denominada «ética de la IA», insistimos en que mejor que de ética, debería hablarse de moral y, sobre todo, de política de la IA. Tres planos que aparecen frecuentemente confundidos, pese a que resulta imprescindible diferenciarlos. Tras criticar a los ingenieros, científicos y filósofos que conciben los sistemas dotados de IA como agentes éticos, perfilamos los principales riesgos que plantea esta tecnología: la amenaza a la privacidad digital; el sesgo en los datos y los algoritmos; la explicabilidad y la responsabilidad por los errores; e implicaciones sociales como la manipulación informativa. Con relación a esta andanada de problemas, analizamos las diferentes normativas para una «IA ética» que la Unión Europea y otras organizaciones están proponiendo, concluyendo que detrás de los sistemas con IA están los intereses de los gigantes tecnológicos y de ciertos Estados, como EE.UU. y China. No es, como ingenuamente cree Noah Harari, que si las decisiones de las máquinas son menos estúpidas que las nuestras, delegaremos en ellas ciertas decisiones, sino que las estaremos delegando a los grupos que están detrás de los sistemas con IA. No es que las máquinas vayan a controlar a los hombres, sino que ciertos hombres van a controlar a otros hombres a través de las máquinas.

Concluimos. Antaño, en el siglo XVII, el padre Malebranche se equivocaba con los animales, queriendo desterrarlos del eje angular y confinarlos al eje radial, al exclamar, mientras oía el lastimero aullido de una perra preñada que era golpeada: «¡Es una máquina! ¡No siente nada!». Hogaño, sin embargo, no se equivocaría con los sistemas con IA, que no están como quieren algunos a punto de poblar el eje angular sino que siguen insertos en el eje radial, acaso cual fetiches, como cosas que parecen tener fuerza por sí mismas. Si el padre Malebranche exclamase, al ver a un hombre dispuesto a desenchufar un robot: «¡No se preocupe! ¡Es una máquina! ¡No siente nada!», pues tendría toda la razón.



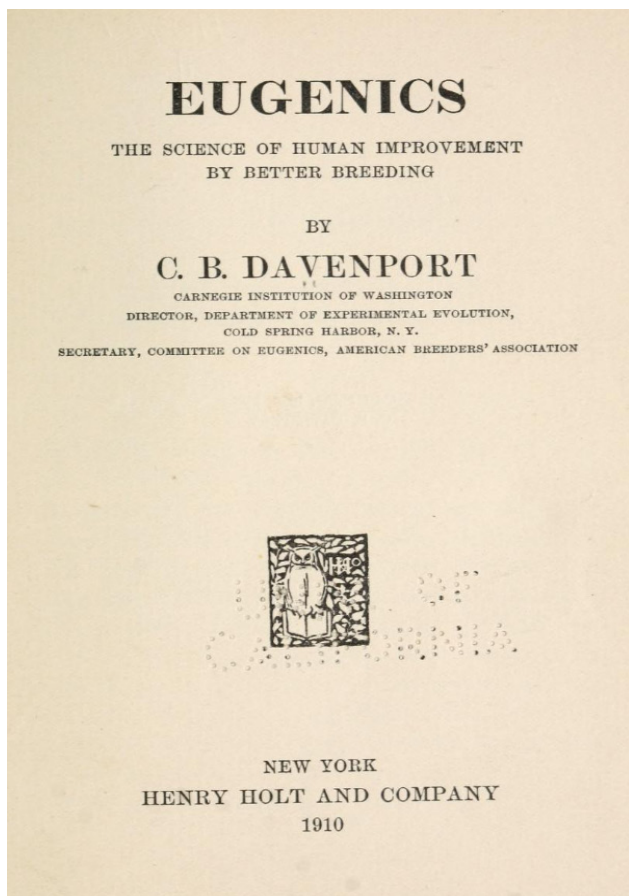
## Coda

### *Transhumanismo, inteligencia artificial y biotecnología*

Como remate del opúsculo queremos analizar el transhumanismo, esto es, la filosofía espontánea de muchos científicos de la computación y biotecnólogos. Esta filosofía o, más bien, dados sus mimbres, nematología surge asociada a una antropotecnia que aspira a una modificación profunda del ser humano por medios tecnológicos *info* y *bio*, persiguiendo aumentar sus rendimientos (resistencia, percepción, cálculo, memoria, longevidad...) y alcanzar, remotamente, la inmortalidad.

Suele definirse el transhumanismo como «la búsqueda tecnológica del mejoramiento humano» (Diéguez 2017). Pero, aunque sus partidarios hablen en ocasiones de «la ciencia del mejoramiento humano», no estamos ante una ciencia sino ante una pluralidad de *info*- y *bio*-tecnologías que, aparte de un momento tecnológico, comportan un momento nematológico exacerbado. A este respecto no está de más acordarse de la vieja definición de la eugenesia como «la ciencia de la mejora de la raza», ofrecida por Francis Galton en 1883, o como (literalmente) «la ciencia del mejoramiento humano», según Charles Davenport en 1910 (Madrid Casado 2015). Una vez apagados los hornos crematorios al terminar la II Guerra Mundial, las razas de la «ciencia eugenésica» fueron declaradas inexistentes por la UNESCO en 1950, y así la «ciencia transhumanista» se centra en la especie humana en

general, cambiando las técnicas de los eugenistas, basadas en el cruce de individuos seleccionados, la esterilización, el aborto o el infanticidio, por nuevas técnicas salidas de la ingeniería genética, la biología sintética, la robótica o la IA.



*Portada del libro de 1910 del eugenista estadounidense Charles Davenport donde la eugenesia es definida como la «ciencia del mejoramiento humano». En lugar de «improvement», los transhumanistas emplean en inglés «enhancement» (así aparece, por ejemplo, en el título del libro editado por Julian Savulescu y Nick Bostrom en 2009), que también significa mejora, pero con el matiz de potenciamiento, como si las cualidades a mejorar ya fueran buenas.*

Para los transhumanistas, el eje angular del espacio antropológico<sup>73</sup> será poblado en un futuro no muy lejano por una nueva especie transhumana, posthumana. Para la vertiente informática del transhumanismo, por el cibernético, una criatura compuesta de partes orgánicas y dispositivos cibernéticos, que fusiona al hombre con la máquina. Y, para la vertiente biotecnológica, por *homo excelsior*, una nueva especie biológica creada a partir de la edición genética de *homo sapiens*. Sin perjuicio de que la mejora de la condición humana mediante las nuevas tecnologías sea uno de los temas bioéticos más discutidos en el presente en marcha, los transhumanistas quieren muchas veces hacer pasar por tecnociencia lo que no es sino una ideología contaminada por una metafísica cuyos compromisos gnoseológicos, ontológicos y políticos hay que sacar a la luz.

Pero comencemos por el principio.

1. *Humanismo*. Es lugar común afirmar que el humanismo renacentista buscó situar al hombre en la cúspide de la *Scala Naturae*. Pero es discutible que los humanistas, es decir, los versados antes en las letras humanas que en las letras divinas, profesaran el humanismo, dado que términos como «humanismo» y «renacimiento» no aparecieron hasta los siglos XVIII y XIX, respectivamente (Bueno 2015). Más bien, el humanismo, definido como «amor general por la humanidad» en el París de 1765, fue consecuencia de una serie de transformaciones —el descubrimiento de América, la Reforma, la Revolución Científica, &c.— que caracterizan a la Modernidad y que determinaron una inversión teológica. Fue en el contexto de esta inversión teológica que el Hombre terminó ocupando el puesto que anteriormente correspondía a Dios como centro del Universo. El humanismo que cristaliza en la segunda mitad del llamado *siglo de*

---

73 Desde los postulados del materialismo filosófico, el espacio antropológico no es bidimensional (Reino de la Naturaleza/Reino de la Gracia, Naturaleza/Cultura, &c.) sino tridimensional, constando de un eje radial (que comprende a las entidades impersonales o cosas), un eje circular (que comprende a las entidades personales antropomorfas, a los hombres) y, atención, un eje angular (que comprende a las entidades personales no antropomorfas, es decir, a los animales, una vez descartados los dioses, los ángeles y los demonios e, incluso, esa Superinteligencia basada en la IA que algunos imaginan como numen o centro de voluntad independiente) (Bueno 1996b, Lectura II).

*las luces* no fue la causa sino el efecto de la inversión teológica. No es casualidad que Linneo definiera al hombre como *homo sapiens*, esto es, mediante la sabiduría, un atributo en origen divino, al igual que la razón o la libertad, facultades convertidas también en características distintivas del ser humano frente a las cosas y los animales (considerados como máquinas por Gómez Pereira y Descartes).

Son varias las ideas-fuerza involucradas en esta permutación radical. Desde la concepción del pensamiento, la razón o la libertad como atributos sustantificados intrínsecos a la naturaleza humana, a la organización climacológica y progresista de la realidad que conlleva la idea de una *Scala Naturae*. Pero, acaso, la principal y más potente sea el mito de la Humanidad, un mito oscuro y confuso de actualidad casi permanente, que consiste en creer que existe el Género humano como ser viviente capaz de ser amado, que existe el Hombre como unidad más allá de su común género biológico («¡Humanos de todos los países, uníos!»), a pesar de que, en cuanto tomamos en cuenta las instituciones sociales e históricas, observamos que no existe el hombre como tal, sino que lo que realmente hay son españoles, franceses, alemanes, ingleses, chinos, indios, &c.

2. *Posthumanismo y transhumanismo*. En las últimas décadas han aparecido dos corrientes que cuestionan el humanismo latente (sea existencialista, marxista o cristiano) y pretenden superarlo de una vez por todas. Por un lado, está el posthumanismo, que lo hace por la vía cultural. Esta es la posición del postmodernismo, del linaje que va de Foucault y ciertos intelectuales franceses a los estudios postcoloniales, el ecologismo, el feminismo, la teoría *queer*, &c. Por otro lado, está el transhumanismo, que intenta superar el humanismo no por la vía cultural sino por la vía científico-tecnológica. Este movimiento, de moda en los últimos tiempos, tiene entre sus apologetas al famoso historiador Yuval Noah Harari, que en *Homo Deus* acaba abogando, precisamente, por el transhumanismo (Noah Harari 2019a).

El movimiento transhumanista promueve, como va dicho, el mejoramiento del ser humano: curar enfermedades hoy incurables, aumentar nuestras capacidades físicas y cognitivas, prolongar la vida humana y, en el límite, crear una especie post-humana o super-humana (*homo excelsior*) que podría llegar a ser inmortal, pues no sólo se trataría

de retrasar el envejecimiento 10, 20, 50 ó 100 años, sino de retrasarlo indefinidamente. Uno de sus gurús, José Luis Cordeiro, profesor de la Universidad de la Singularidad de Silicon Valley (financiada por Google), afirma sin rubor en las entrevistas que el primer ser humano inmortal, que no morirá de viejo, ya ha nacido y habita entre nosotros (no sé, lector, si tú serás uno de ellos). A su juicio, el envejecimiento será una enfermedad curable mediante fármacos y otras terapias a partir de 2045, de manera que el ser humano nunca más morirá de viejo sino, a lo sumo, como consecuencia de un accidente, de que le atropelle un tráiler o le caiga un piano de cola en la cabeza. De nuevo, este embeleso recuerda al que Francis Galton imaginara en la novela utópica que escribió hacia el final de su vida, titulada *Kantsaywhere (Donde dijo Kant)*, en la que una sociedad que vivía bajo los preceptos eugenésicos dictados por sacerdotes-científicos progresaba procreando seres cada vez mejor dotados (por cierto, que la sobrina y heredera de Galton, dado que éste curiosamente no tuvo descendencia en su matrimonio, quemó parcialmente el manuscrito irritada por algunas escenas reproductivas subidas de tono). Como puede comprobarse, el papel lo aguanta casi todo, desde los sueños de Galton a los de Cordeiro o Noah Harari.

Es lugar común cifrar el origen del movimiento transhumanista en el artículo que el filósofo Max O'Connor –hoy rebautizado sintomáticamente como Max More– publicó con el título «Transhumanismo: hacia una filosofía futurista» en 1990 (Diéguez 2017, cap. 1). No obstante, el rótulo «transhumanismo» fue acuñado por Julian Huxley, biólogo evolutivo y eugenista británico, en su libro *Religión sin revelación* de 1927, para referirse a la posibilidad que la especie humana tiene de trascenderse a sí misma. Sin embargo, para encontrar un significado similar al actual, como transformación del ser humano mediante la ciencia y la tecnología, hay que fijarse en otras obras de la misma década debidas a J. B. S. Haldane o John D. Bernal, así como en la célebre novela *Un mundo feliz* de 1932 de Aldous Huxley, pese a que en ellas no aparece el término como tal.

En su artículo de 1990, O'Connor/More definía el transhumanismo como el conjunto de filosofías que buscan guiarnos hacia una condición posthumana. Reconocía que el transhumanismo comparte con el humanismo el respeto por la razón y el compromiso con el progreso; pero señalaba que difería en fines y medios, pues el transhumanismo

aspira a cambiar, no ya las condiciones de la existencia humana (mediante la educación y la cultura), sino la propia naturaleza humana (su herencia biológica y genética, que no contempla como perfecta ni intocable), basándose en la neurociencia, la farmacología, la biotecnología, la nanotecnología, la ultrainteligencia artificial, &c.

Se debe a la mujer de O'Connor/More, la artista nacida como Nancie Clark y conocida como Natasha Vita-More (*sic*), el *Manifiesto Transhumano*, difundido en 1983 y hoy refundido en la *Declaración Transhumanista* de 2009, firmada por ella y otra veintena de autores (cuya traducción al español puede leerse en el Anexo IV). Entre los ocho puntos de la declaración hacen acto de presencia las ideas-fuerza referidas a la razón, el progreso y, especialmente, la Humanidad, concebida como una clase dotada de una unidad e identidad que desbordan el marco meramente biológico. Y en el séptimo punto puede leerse:

Abogamos por el bienestar de todo ser sensible, lo que incluye a los humanos, los animales no humanos y cualesquiera intelectos artificiales futuros, formas de vida modificadas u otras inteligencias a las que pueda dar lugar el avance tecnológico y científico.

Actualmente, el tecnomejoramiento humano es materia de investigación en varios centros punteros de la angloesfera: el Instituto para el Futuro de la Humanidad, liderado por Nick Bostrom y perteneciente a la Universidad de Oxford; el Centro Uehiro para la Ética Práctica, dirigido por Julian Savulescu y también perteneciente a la Universidad de Oxford; y, por no seguir, la Universidad de la Singularidad (por la «singularidad tecnológica»), conducida por Ray Kurzweil y sita en Silicon Valley, con el patrocinio de Google o la NASA. Por su parte, la Asociación Mundial Transhumanista, fundada en 1998 por Nick Bostrom y David Pearce, ha cambiado su nombre a Humanity Plus (H+) y cuenta con más de cinco mil miembros en todo el mundo.

3. *Transhumanismo e inteligencia artificial*. El transhumanismo tiene dos ramas: la versión *info* o computacional y la versión *bio* o biotecnológica. La primera modulación enlaza con la investigación en IA, contando con apóstoles como Kurzweil (2012) y Bostrom (2016),

y aspira a que la especie transhumana sea el *cíborg* (término que proviene de la contracción de «*cibernético organismo*»), una suerte de hombre-máquina, en el que los órganos biológicos se han fusionado con implantes robóticos (prótesis de *hardware* o *software*), responsables de la mejora de las capacidades físicas y cognitivas de este nuevo ser. La integración total con la máquina se conseguiría desarrollando interfaces permanentes cerebro-ordenador. Este enlace íntimo con la máquina –por decirlo con Kurzweil (2012, 33)– es nuestra mejor opción para hacer frente a la superinteligencia artificial que, supuestamente, llegará como consecuencia de la singularidad tecnológica y excederá al intelecto humano en prácticamente todas las áreas de interés (Bostrom 2016, 22). Adelantándose al futuro, algunos, como Chris Gray, incluso plantearon en 1997 una carta de derechos del cíborg, postulada como enmienda a la constitución de los EE.UU., donde persiguen el reconocimiento para los cíborgs de derechos como el de libertad de movimiento, la libertad de expresión electrónica, la privacidad electrónica, la libertad de familia, sexualidad y género, y el derecho a la paz.

Además, los transhumanistas de esta rama especulan con que la inmortalidad se podría conseguir mediante el volcado de nuestra memoria en un ordenador y su posterior transferencia a un robot (*mind uploading* y *mind transfer*). Los cerebros humanos dejarán paso a computadoras electrónicas más potentes que incorporarán nuestros recuerdos y sensaciones. Así, el desarrollador de robots y tecno-entusiasta Hans P. Moravec (1999) profetiza que los humanos transferiremos nuestras mentes a robots artificialmente inteligentes, que serán la progenie mecánica de la especie humana.

Pero, ¿en qué se basan para realizar afirmaciones tan aventuradas? ¿Cuál es su fulcro de verdad? En que hoy día los humanos disponemos de varias prótesis que funcionan con éxito, como los marcapasos, los implantes cocleares e, incluso, los implantes de brazos robóticos (un hito logrado en 2013). Es más, el ingeniero y profesor de cibernética Kevin Warwick ha trabajado en el implante de un cerebro biológico, hecho con neuronas, en un robot, a fin de dirigirlo; y, recíprocamente, en el implante de electrodos de estimulación cerebral profunda para vencer los efectos de la enfermedad de Parkinson (Warwick 2012). El implante de otras interfaces cerebro-ordenador ha servido para que los pacientes muevan un cursor en una pantalla, controlen una mano

robótica o conduzcan una silla de ruedas. Recientemente, en agosto de 2023, investigadores de las Universidades de California y Berkeley han dado a conocer la interfaz cerebro-ordenador que ha permitido volver a hablar a una mujer con parálisis provocada por un ictus (un concienzudo entrenamiento durante semanas ha servido para que la IA asocie patrones de actividad cerebral con fonemas y sonidos). Todos estos avances serían (supuestamente) pasos en la dirección del cibernético.

Ahora bien, dejando aparte que las interfaces cerebro-ordenador actuales son limitadas y que la idea de descargar digitalmente toda nuestra mente en una máquina resulta ridícula, hay que hacer una crítica radical a esta corriente del transhumanismo puesto que está completamente envuelta en el mito de la IA fuerte y general, que hemos criticado a lo largo del opúsculo.

Para los transhumanistas computacionales, no es que haya una analogía entre el cerebro humano y el ordenador (algunas características comunes), sino que un cerebro es lo mismo que un ordenador y viceversa, lo que posibilitaría integrar definitivamente uno en otro. El cerebro sería un ordenador digital *sui generis*, constituyendo la mente su programa, un programa trasladable a la CPU de un robot. Pero un cerebro y un ordenador no son, punto por punto, equivalentes. No es sólo que el material de que están hechos no sea el mismo, sino que los sistemas dotados de IA carecen formalmente de comprensión semántica (*apotética*), al reducirse todo en ellos a procesos sintácticos (*paratéticos*), como estudiamos en el caso de ChatGPT páginas atrás (los significantes que maneja esta red neuronal gigante no refieren a las cosas, como en el lenguaje humano, donde se da una circularidad palabra-cuerpo, sino a otros significantes, en función de la frecuencia con que aparecen concatenados, determinando un lenguaje-máquina desencarnado). La inteligencia de los sistemas con IA es deductiva e inductiva; pero no, abductiva (véase el capítulo 3). Estos sistemas generan mayormente relaciones formales (deducciones, predicciones o patrones estadísticos) al margen de los materiales y las conexiones causales realmente existentes, pues se trata –por decirlo en términos teológicos– de «inteligencias separadas».

Además, en lo que toca, no ya al término «inteligencia», sino al término «artificial» conformante del rótulo propagandístico «inteligencia artificial», los transhumanistas computacionales



permanecen presos del mito maniqueo Naturaleza/Cultura. Ante la vieja inteligencia natural propia del hombre, se levantaría la nueva inteligencia artificial de los robots. Pero resulta que la inteligencia humana no reside en los genes y es –según defendimos en el capítulo 4– también artificial, en tanto en cuanto depende de la cultura objetiva, de ingenios objetivos, esto es, de artificios como la escritura, los libros, los laboratorios, la educación, la lengua, la cultura y toda una larga serie de instituciones. Una lista en la que hay que incluir a los mismos ordenadores, cuya inteligencia –recíprocamente– no está encerrada en su carcasa con independencia de los programadores y los ingenieros. Porque no es el ordenador el que explica al cerebro humano, sino que es al revés, es el cerebro humano –o más bien un conjunto de cerebros, un grupo de hombres (pues hasta los periféricos de un ordenador están a la escala quirúrgica de los cuerpos humanos)– el que explica al ordenador y su estructura. La IA sin la IH («inteligencia humana») no es nada, puesto que somos nosotros los que suministramos al sistema el algoritmo y los datos y, en último término, los que interpretamos los estados físicos de las máquinas como movimientos de ajedrez, cálculos aritméticos, estimaciones estadísticas o respuestas a preguntas, dotándoles de contenido semántico y finalidad propositiva. En verdad, aunque esté de moda hablar de aprendizaje de las máquinas (*machine learning*), somos los humanos los que aprendemos, con la asistencia de las máquinas.

Los seres humanos ampliamos y externalizamos nuestra inteligencia con ayuda de las máquinas («inteligencia extendida»). Es lo que en su momento expusimos de que la racionalidad subjetual y la racionalidad objetual se dan siempre conjugadas (véase el capítulo 4). Por esta razón, concluimos, los cibernéticos seguirían perteneciendo al eje circular antes que al eje angular; pues el hombre, más que un mono desnudo, es un mono vestido, vestido con ropajes y equipado con útiles de muy diversa índole (desde un hacha de sílex a un iPhone con auriculares inalámbricos).

4. *Transhumanismo y biotecnología*. La rama más pujante del transhumanismo es, no obstante, la rama biotecnológica, que apuesta por la ingeniería genética y la biología sintética para rediseñar el ser humano. Dentro de esta rama también aparece el biomejoramiento

químico, que estudia la fabricación de fármacos que aumenten de forma explosiva nuestras capacidades físicas o mentales. De hecho, ya existen drogas en esta línea, pastillas o dosis que cambian el volumen de nuestras fuerzas y ánimos, a la manera que la inyección de oxitocina mejora la empatía y la sensibilidad en los varones (Lara & Savulescu 2021, VIII) y, por ejemplo, se está probando la estrategia de rejuvenecimiento físico y cognitivo basada en la inyección de la proteína PF4, que funciona como factor plaquetario que supuestamente otorga mayor plasticidad cerebral.

Pero el camino más directo para el mejoramiento humano pasa por la modificación de nuestros genes en embriones humanos, o sea, afectando a la línea germinal o reproductora (a las células germinales, no sólo a las células somáticas). Una línea prometedora desde que hace poco más de una década se dispone de la tecnología de edición genética CRISPR-Cas9, en cuya investigación fue clave el trabajo pionero del español Francis Mojica, que reparó en la existencia de bacterias transgénicas, que cambian su propio ADN para hacerse resistentes a virus. Este sistema, extraído de ciertas bacterias, es capaz de trabajar como unas tijeras moleculares del genoma. CRISPR-Cas9 es una herramienta más segura a la hora de cortar, insertar, eliminar y reorganizar el ADN que otras disponibles hasta la fecha (ZNF, TALEN). Mientras que la ingeniería genética clásica (la de la oveja Dolly) trabajaba con nuclóvulos (óvulos desnucleados a los que se inserta un ADN completo, operando una transferencia de núcleos, con fines tisulares –crear tejidos para curar al donante– o con fines reproductores), esta nueva tecnología permite editar o corregir genes concretos durante la fecundación, lo que abre la puerta a la obtención de organismos genéticamente modificados a la carta. Así, el debate bioético ha pasado de discutir la clonación de individuos a debatir la eliminación de genes perjudiciales, la potenciación de genes beneficiosos e, incluso, la inserción de genes de otras especies o sintetizados artificialmente en el laboratorio.

Tanto en gusanos como en moscas se han identificado varios genes relacionados con el envejecimiento, cuya modificación ha conseguido alargar su vida significativamente (en el caso del gusano nematodo, se ha logrado multiplicar la duración media de su vida por más de seis). También se ha conseguido alargar significativamente la vida en

ratones, así como diseñar super-ratones resistentes al cansancio y la fatiga. Y, en 2018, el científico chino He Jiankui modificó mediante CRISPR-Cas9 los genes de dos embriones humanos gemelos, provenientes de una pareja en la que el padre era seropositivo, para intentar hacerlos resistentes al VIH. Aunque en China, como en la mayor parte del mundo, está prohibido implantar los embriones resultantes por los riesgos que conlleva (cambios indeseados en el genoma consecuencia de la edición cuyo efecto se desconoce), el científico chino y su equipo lo hicieron. Pero las dos niñas nacidas presentan mosaïcismo: sus células poseen dos códigos genéticos diferentes (uno modificado y otro no). El experimento suscitó la condena unánime de la «comunidad científica», no tanto por aplicar la edición genética con fines reproductivos, cuanto por correr un riesgo notable a sabiendas de que la probabilidad de que un hijo nazca con el VIH por transmisión paterna es bastante baja.

Aunque no es lo mismo operar con bacterias, moscas o ratones que con seres humanos, donde la complejidad es mucho mayor, la aplicación de CRISPR-Cas9 podría servir, por ejemplo, para reparar uno de los genes causantes de la miocardiopatía hipertrófica, entre otras enfermedades. De hecho, en diciembre de 2023, la Agencia Europea del Medicamento ha dado el visto bueno a la primera terapia de edición génica con CRISPR-Cas9 indicada para el tratamiento de dos enfermedades graves de la sangre. Es el primer paso, según los transhumanistas biotecnológicos, hacia esa especie posthumana cuya esperanza de vida sería superior a los 500 años, con capacidades físicas, emocionales e intelectuales mayores a las que el hombre actual tiene (no obstante, como no hay una correspondencia biyectiva entre genes y rasgos, editar un gen puede ocasionar cambios imprevisibles aparte del buscado, lo que siembra la duda acerca de la edición genética como herramienta transhumanista).

A continuación, vamos a pasar revista a los argumentos filosóficos en contra (a) y en pro (b) del transhumanismo en su versión biotecnológica que suelen ofrecerse:

a.1) *Contraargumento de la naturaleza humana.*

El primer argumento en contra del biomejoramiento genético es el que atañe a la preservación de la naturaleza humana. Entre los llamados

«bioconservadores», enfrentados a los llamados «bioprogresistas», encontramos a Fukuyama (2003) y Habermas (2002). De hecho, Fukuyama (2004) tilda al transhumanismo de la idea más peligrosa del mundo.

Estos críticos sostienen que existe algo así como la naturaleza humana y que no estaría bien violarla o modificarla. El emparejamiento selectivo ha implementado cierta selección genética en la especie humana; pero la ingeniería genética a la carta concentraría esta selección en cierta clase social (los que pudieran pagarla), esquivando ese mecanismo igualitario que es la lotería genética, donde el resultado es en principio igual de justo o injusto para todos, por encima de la clase social a la que cada uno pertenezca (o, al menos, el desconocimiento de su desenvolvitura procura la apariencia de igualdad a todos los participantes: la madre no elige, de momento, al padre en función de su dote genética).

Como puede constatarse, el transhumanismo pone sobre la mesa el candente tema filosófico de la naturaleza humana; porque si se quiere ir más allá de lo humano, hay que preguntarse qué es el ser humano. Y aquí caben varias opciones, que podemos organizar en función del peso (0 ó 1) que se otorgue a la Naturaleza (N) y la Cultura (C) en la determinación de qué es el ser humano:

Naturaleza	Cultura	Proponentes
1	0	Fukuyama, Habermas
0	1	Ortega, Diéguez
1	1	Aristóteles, Marcos
0	0	Materialismo filosófico

- La primera alternativa es la que afirma la existencia de una naturaleza humana a la que la cultura habría de subordinarse:  $(N, C) = (1, 0)$ . Esta es la posición de Fukuyama (no tanto de Habermas). La naturaleza humana no residiría, como antaño, en el alma racional, sino ahora en el código genético.

El problema es que nuestro ADN no es común a todos los miembros de la especie y, además, es compartido en gran medida por otros primates. Es más, el ADN humano cambia en el transcurso

de la evolución, por lo que difícilmente puede ser el sustento de una naturaleza fija e inmutable. Por otro lado, tan natural es un estado saludable como un estado enfermo. Nada más natural que una apendicitis y, sin embargo, actuamos para corregir el curso natural de los acontecimientos. Para el biólogo, un tumor puede ser un interesante caso de estudio producto natural de una mutación, y, sin embargo, para el médico, por más natural que sea, no deja de ser una aberración –un regalo envenenado de la naturaleza– a extirpar, buscando curar al afectado.

- La segunda alternativa es la que niega la existencia de una naturaleza humana y todo lo fía a la cultura:  $(N, C) = (0, 1)$ . Esta es la posición de Ortega («el hombre no tiene naturaleza, lo que tiene es historia»), para el que el ser del hombre es constitutivamente movilidad y cambio, reacción enérgica contra la naturaleza, que le lleva a crear una sobrenaturaleza artificial mediante la técnica.
- La tercera alternativa es la que alcanza un compromiso entre ambos polos, entre Naturaleza y Cultura:  $(N, C) = (1, 1)$ . Así, para Aristóteles, el hombre era el animal social («zoon politikón»).

En España, hay planteado un interesante debate entre la segunda y la tercera alternativas, que es el que sostienen, por un lado, Alfredo Marcos (2018), que reivindica a Aristóteles, y, por otro lado, Antonio Diéguez (2017, 138-139), que sostiene, haciendo suya la interpretación de Ortega de José Lasaga (2003), que, efectivamente, el hombre lo que tiene es historia, no tiene naturaleza, más que como una especie de equipaje que estaría en la mochila. En consecuencia, para Diéguez (2017, 194), hay que «evitar el error común de realizar juicios generales y definitivos, de lanzar condenas o alabanzas globales». Se trataría de evaluar cada propuesta transhumanista por separado, no en bloque, atendiendo a sus consecuencias, dado que no existe una naturaleza humana que tomar como principio y criterio. Marcos (2018) sustenta que la filosofía de Ortega reivindicada por Diéguez (2017) resulta inoperante en el contexto de la discusión del transhumanismo, al estar lastrada por el antidarwinismo de Ortega, así como que Diéguez interpreta la idea de naturaleza humana

en términos platónicos, no aristotélicos, como idea atemporal y universal, y por eso la desecha. Pero Marcos (2018, 121-122), a la hora de construir una idea no platonizante de naturaleza humana, con implicaciones normativas, se desliza a la metafísica, al incluir rasgos animales y sociales pero también espirituales, desembocando *velis nolis* en lo religioso.

— Por último, la cuarta alternativa —representada en la matriz como  $(N, C) = (0,0)$ — se sitúa más allá del reduccionismo natural o cultural, así como de su yuxtaposición ecléctica. Frente a los esquemas *metaméricos* que tienden a reducir un polo a otro, o bien a yuxtaponerlos, el materialismo filosófico ensaya un esquema *diamérico* de conjugación, donde Naturaleza (N) y Cultura (C) se nos muestran como multiplicidades ( $N_1, N_2, N_3\dots$  y  $C_1, C_2, C_3\dots$ ) cuyas partes cabe conjugar, por ejemplo en el momento de establecer un canon de individuo humano ( $N_1 - C_2 - N_3 - C_4 - \dots$ ). Una ilustración de la conjugación o el entreveramiento entre el plano natural y el plano cultural, de su retroalimentación circular, nos la proporciona la detección de un incremento en la prevalencia de la sordera congénita, explicable por lo que algunos genetistas llaman coevolución, y nosotros llamamos codeterminación, entre biología y sociedad. Los sordos de nacimiento pueden hoy día hablar mediante sus lenguajes de signos, lo que hace más probable que se casen entre ellos, fomentando la supervivencia del gen frente a la presión de la selección natural. La rectificación de las ideas metafísicas de Naturaleza y Cultura se opera mostrando la trabazón entre componentes biológicos, etológicos, antropológicos e históricos.

Se trata, en suma, de no caer en el mito de la Naturaleza pero tampoco en el mito de la Cultura (Bueno 2016), huyendo de la visión dualista de la Naturaleza y la Cultura como dos todos inconmensurables, así como de la concepción binaria del hombre como una suerte de centauro ontológico (por decirlo otra vez con Ortega). Naturaleza y Cultura no conforman una distinción disyuntiva, existiendo configuraciones —como va a ocurrir con el canon de individuo humano— que no cabe juzgar, si no nos sentimos

aprisionados por la dicotomía, como naturales ni como culturales, sino como de una tercera clase.<sup>74</sup>

El materialismo filosófico sustenta que, efectivamente, hay un canon humano; pero lo que ocurre es que ese canon no es una esencia porfiriana –condiciones necesarias conjuntamente suficientes– sino plotiniana, esto es, una esencia procesual, que va cambiando con el tiempo, como va cambiando la altura media de los españoles (en la Edad Media, un varón de la península ibérica medía 1,60 metros de media, lo que contrasta con la estatura media actual en España, que está en 1,76 metros). Navegando entre Escila y Caribdis, hay que evitar recaer en una ontología fijista y sustancialista, sin por ello negar de plano la ontología de la esencia humana. Se trata de alcanzar una visión dinámica, dialéctica, que es lo que el materialismo filosófico persigue al introducir la idea del canon de individuo corpóreo operatorio humano (que define precisamente la norma que los médicos intentan conservar con su praxis).

Más allá del reduccionismo biologicista (el hombre tiene una naturaleza predeterminada) o sociologista (el hombre es mera construcción social), y de su yuxtaposición ecléctica, el canon humano se nos aparece como «híbrido» que conjuga aspectos naturales y culturales, que está delimitado biológica e históricamente, desbordando por ello la oposición metafísica entre Naturaleza y Cultura comprendidas como todos enterizos y estancos (Madrid Casado 2022b). Este canon define una morfología elemental de la individualidad corpórea humana (canonizada, precisamente, por Policleto). Pero este canon nos es dado *in medias res*, en medio de la evolución natural y del proceso histórico: no basta con apelar a haber nacido de otro humano o a pertenecer a la especie humana por tener tal genoma, ya que el canon desborda la biología, dado que el hombre es un animal institucional, cuya morfología no está únicamente escrita naturalmente sino también transformada socialmente, pues –por ejemplo– «el fuego hizo al hombre» (o piénsese en cómo el Doríforo, el canon de Policleto, portaba una lanza). La morfología canónica

---

74 A la manera de las entidades científicas, el canon constituye una hiperrealidad, una estructura que no es descubierta naturalmente pero tampoco inventada socialmente, sino que se trata –conjugando dialécticamente ambos extremos– de un descubrimiento constitutivo (Madrid Casado 2018a, cap. 17).

del individuo humano en cuanto sujeto corpóreo operatorio no está predeterminada por una sustancia estática que actuase debajo del canon, sino que está codeterminada procesualmente por el conjunto de individuos de una clase (especie o grupo) en evolución biológica y desarrollo histórico (Bueno 2001, 104-106).

a.2) *Contraargumento de la dignidad humana.*

El segundo argumento que se da en contra de las modificaciones genéticas que promueve el transhumanismo aduce que éstas alterarían, no ya la naturaleza humana (rechazada su existencia), sino la dignidad humana. Así, para Habermas (2002), la ingeniería genética no atendería tanto contra una supuesta naturaleza humana fija e inmutable sino contra una supuesta dignidad humana, cuya indisponibilidad subraya el filósofo alemán, al ser imposible para los padres saber qué don genético consensuar con el hijo con anterioridad al nacimiento. Por decirlo de otra manera: si seleccionan su identidad genética, los padres están de facto violentando la dignidad del hijo, al coartar su libertad, basada en la recombinación y la lotería genética. Los padres que predispusieran genéticamente a su hijo serían responsables de los efectos ocasionados por la manipulación. Mientras que en el caso de los padres que no lo hacen, la configuración genética del hijo no es responsabilidad de nadie, salvo quizá del mecanismo biológico. Sin embargo, hay que advertir que prácticamente nadie vería con malos ojos la corrección de ciertas enfermedades hereditarias graves en la descendencia en caso de que fuera factible.

No deja de ser paradójico que el argumento de Habermas puede retorcerse para condenar el aborto (algo de lo que el propio Habermas se cuida), porque ¿acaso los padres están en condiciones de consensuar con el futuro hijo la técnica abortiva a aplicarle?

Pero el lastre principal del argumento es que pretende explicar lo oscuro por lo más oscuro. ¿De qué hablamos cuando hablamos de «dignidad humana»? Pico della Mirandola y, posteriormente, Kant definieron la dignidad humana en función de la libertad o la moralidad: las cosas tienen un precio, un valor relativo, mas las personas no, siendo irremplazables unas por otras. Pero estamos ante una idea metafísica, trasunto secularizado de la gracia divina como consecuencia de la inversión teológica. La dignidad del hombre es



defendida, en el plano de la representación, frente a los animales y los ángeles; porque el hombre tiene cuerpo –como los animales– y espíritu –como los ángeles–. En el fondo, la dignidad humana proviene de que el hombre recibe la gracia (santificante) de Dios y, especialmente, de que Dios se encarnó en Jesucristo, asumiendo naturaleza humana en el vientre de la Virgen María. La dignidad del hombre es, en origen, la dignidad del hombre cristiano, defendida, en el plano del ejercicio, frente a los judíos y los musulmanes. En palabras de Bueno (2015, 2):

La «dignidad del Hombre», exaltada por escritores renacentistas como Pérez de Oliva o Cervantes de Salazar, significa principalmente, en el contexto histórico, que el hombre está por encima de los ángeles, puesto que Jesucristo se encarnó en un animal, y no en un querubín; por lo que la dignidad del Hombre equivalía, indirectamente, a una expresión de la conciencia de superioridad de los cristianos frente a los musulmanes, que consideraban el dogma de la divinidad de Cristo como una blasfemia.

Para rizar el rizo, el transhumanista Nick Bostrom asevera hablar «en defensa de la dignidad posthumana» (Hansell & Grassie 2010, cap. 2). Lo cual es como si habláramos en nombre de la dignidad angelical o de la dignidad de los extraterrestres (no en vano, existe cierta analogía *emic* entre el cibernético u *homo excelsior* y el ángel o los marcianos, en cuanto a su carácter pretendidamente angular): una impostura.

### a.3) *Contraargumento de las desigualdades económicas.*

El tercer y último argumento que suele ofrecerse en contra del transhumanismo es el que apunta que las modificaciones genéticas sólo las podrían pagar los ricos y no los pobres, lo que generaría mucha desigualdad. Pero, claro, esto no es algo nuevo ni inaudito, pues igual ocurre con la mayoría de innovaciones tecnológicas.

### b.1) *Argumento de la mejora por la mejora.*

Pasemos ahora a desbrozar los argumentos que se dan a favor del transhumanismo. El primer argumento en pro del transhumanismo se resume en una pregunta: ¿quién no quiere mejorar? ¿Por qué cerrar la puerta a las mejoras físicas, cognitivas y emocionales que arroja el progreso científico y tecnológico? Es el controvertido imperativo de la mejora.

Lo primero que hemos de preguntar es si realmente son, consideradas en conjunto, mejoras. En otras palabras: si cada uno de los avances científicos o tecnológicos que esgrimen los transhumanistas supone *ipso facto* una mejora. Más bien, habrá que ir analizando caso por caso, discriminando los pros y los contras de cada intervención que se postule. De lo contrario, lo que estará funcionando a toda máquina es el mito del progreso.

Una distinción a la que se suele acudir al reflexionar sobre este tema es que, en las intervenciones (cibernéticas o genéticas) que se puedan hacer, hay que distinguir dos tipos: las que son terapéuticas, para curar o prevenir; y las que son meliorativas, para mejorar.<sup>75</sup> Por ejemplo, cuando el congreso transhumanista TransVision se celebró en octubre de 2021 en el Colegio de Médicos de Madrid, con la presencia de Ray Kurzweil y José Luis Cordeiro, muchos médicos protestaron porque los temas que iban a abordarse no encajaban con la bioética médica, que no busca mejorar sino curar. La controversia radicaría, presuntamente, en que mejorar es problemático, mientras que curar no lo es.

Pero la distinción entre terapia y mejora, pese a lo extendida que está (Moreno & al. 2015, 60-61), es muy permeable y borrosa, si no se dan parámetros sistemáticos. Por de pronto, transhumanistas como Julian Savulescu se amparan en que las intervenciones que proponen mejoran el bienestar y, por tanto, de acuerdo con la definición de salud de la Organización Mundial de la Salud («la salud es un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades»), serían propiamente terapéuticas. De hecho, Savulescu defiende que «una mejora es un cambio en nuestra biología y psicología que aumenta nuestro bienestar» (Lara & Savulescu 2021, 29).

Además, la distinción terapia/mejora se parece mucho a la que hacían los eugenistas de principios de siglo XX entre eugenesia negativa y positiva (Madrid Casado 2015).<sup>76</sup> La «eugenesia positiva»

---

75 También puede haber intervenciones que no sean terapéuticas ni meliorativas, sino perjudiciales (si lo que hacemos es introducir un gen que daña a la persona) o neutras (si el gen no ocasiona efecto alguno).

76 Es llamativo que la Asociación Transhumanista Mundial, hoy Humanity+, en su web, dentro de las FAQs, busque trazar una línea divisoria con el movimiento eugenésico, al indicar que ellos no pretenden imponer involuntariamente nada porque son tolerantes.

perseguía la reproducción de los individuos más aptos, a fin de incrementar sus características (altura, belleza, inteligencia...). Por contra, la «eugenesia negativa» consistía en evitar la reproducción de los individuos considerados menos aptos, con características indeseables. Esta se veía y se ve generalmente bien; porque, por ejemplo, hoy día, se tiende a evitar que las mujeres con síndrome de Down tengan descendencia o se permite el aborto de embriones o fetos mal formados, con algún tipo de síndrome o enfermedad grave, como consecuencia de los diagnósticos prenatales llevados a cabo mediante técnicas como el diagnóstico genético preimplantacional o la amniocentesis. En estos casos de eugenesia negativa, se suele decir que no se está buscando una mejora, sino que se trata simplemente de una práctica terapéutica.

La dificultad estriba en que, si consideremos la especie humana al nivel de la clase y no del individuo, estas prácticas terapéuticas sí comportan una mejora de la raza. Estadísticamente, si en un conjunto de datos eliminamos los valores inferiores (para entendernos, los individuos menos aptos), la media de los datos inmediatamente cambia, aumenta, al haberse corrido la distribución hacia la derecha, hacia los valores superiores (los individuos más aptos). De resultas, se produce exactamente el mismo efecto —el aumento de la media de los datos— que si, en vez de eliminar los valores inferiores, hubiéramos incrementado los valores superiores sumándoles cierta cantidad. Por consiguiente, toda eugenesia negativa es ya una eugenesia positiva (como en aritmética, menos de menos es más). O, con otras palabras, toda intervención X, aunque a la escala molecular del individuo se defina como terapéutica, resulta meliorativa a la escala molar de la clase, puesto que mejora algunos estadísticos descriptivos del conjunto como la media. En este sentido, la medicina opera *in recto* a escala molecular, pues va dirigida a la salvaguarda de los individuos tomados distributivamente; pero sus acciones terapéuticas implican una mejora *in oblicuo* a escala molar, cuando tomamos a los individuos atributivamente, conformando grupos y sociedades políticas. Como ejemplifica Diéguez (2017, 95-96) para poner de relieve lo poroso de la distinción entre terapia y mejora, una técnica terapéutica como la vacunación implica mejoras en la calidad de vida y, recíprocamente, mejoras en la calidad de vida de los ancianos son

consideradas habitualmente como terapéuticas. Es más, los hábitos de vida dictaminados por la medicina higienista muchas veces persiguen, a pesar de venir prescritos por médicos (por los terapeutas por excelencia), más la mejora que la curación, alargando la esperanza de vida como consecuencia de llevar unos hábitos de vida saludables (práctica regular de ejercicio o deporte, reducción del consumo de alcohol, tabaco y grasas saturadas, &c.).

En este punto, constatada la indefinición de la distinción entre terapia y mejora si no se dan parámetros sistemáticos, parece interesante introducir una distinción escolástica que el materialismo filosófico reivindica –la distinción entre lo privativo y lo negativo– con el objeto de reformular la distinción entre lo curativo y lo meliorativo borrando cualquier adherencia metafísica. Ahora bien, lo hacemos tomando como parámetro de la misma –y esta es la clave del asunto– el canon de individuo humano que mencionamos más arriba, un canon procesual del sujeto corpóreo operatorio que varía biológica e históricamente, pero que los médicos con su praxis se esfuerzan en conservar (Madrid Casado 2022b).

Valga el siguiente ejemplo para explicar la distinción negativo/privativo: si a una persona le falta una tercera pierna, esta ausencia es *negativa*, porque no la tiene pero tampoco debería tenerla, pues –conforme al canon– no es propio de un ser humano poseer tres piernas. Sin embargo, si una persona nace con una sola pierna, la falta de una segunda pierna es *privativa*, porque carece o está privado de algo que debería tener, según el canon. Del mismo modo, cuando un bebe nace con seis dedos en una mano, los médicos le extirpan el sexto dedo, buscando ajustarlo al canon, ya que el neonato está *privado* de la pentadactilia humana.

Conviene reparar en que la distinción privativo/negativo no es un calco de la distinción terapéutico/meliorativo. Por un lado, mientras que esta última permanecía en una grosera indefinición entre la escala molecular (el individuo como mónada) y la escala molar (la clase o el conjunto como un todo), la referencia al canon de individuo humano posibilita –por así decir– introducir una relación de equivalencia, basada en la mayor o menor semejanza con el canon, que divide la clase o el conjunto y reorganiza a los individuos en clases de equivalencia (partes formales).

Por otro lado, para mostrar que no hay una correspondencia entre lo terapéutico y lo privativo, así como entre lo meliorativo y lo negativo, fijémonos en un caso real: Sharon y Candy, una pareja de lesbianas sordas estadounidenses que en 2002 quisieron y lograron tener un bebé sordo, usando para ello el semen de otro donante sordo. Para ellas, la tecnología subyacente era meliorativa, un avance o mejora de la ciencia, al permitir su reproducción en el seno de la comunidad sorda. De hecho, de este modo lo vio Julian Savulescu, que pasó de defender la procreación beneficiosa de niños videntes y oyentes mejor que de niños ciegos o sordos mediante la utilización de test genéticos preimplantacionales en la reproducción asistida, pues las parejas tienen la obligación moral de traer mejores hijos seleccionando genes no patológicos (Savulescu 2001), a defender una eugenesia liberal, donde la sordera –como el sexo o el color de ojos– puede ser buscada y seleccionada si no reduce el bienestar (Savulescu 2002). Aunque las progenitoras o el propio Savulescu no consideren la sordera como una discapacidad sino, según se dice ahora, como una diversidad funcional, propia de una comunidad cultural (las personas sordas), desde nuestras coordenadas, que toman como referencia inexcusable el canon procesual de individuo corpóreo humano, la sordera no es –por más que resulte políticamente incorrecto– negativa sino privativa. Los niños nacidos sordos carecen del sentido de la audición, característico de los seres humanos (pero esta diferencia ontológica no tiene necesariamente por qué comportar diferenciaciones éticas o morales, ni por supuesto jurídicas).

Si tomamos como referencia el canon de individuo humano (un canon, empero, que no es natural, al estar mediatizado por la cultura, al ser –por así decir– tan genético como epigenético), podemos distinguir entre lo que son terapias y/o mejoras admisibles dentro de la variabilidad intrínseca del canon (proporcionar una pierna protésica, acoplar un marcapasos, insertar una interfaz cerebro-ordenador que posibilita hablar a un paciente con parálisis, suministrar hormonas del crecimiento a un niño con problemas de crecimiento, corregir la miopía, realizar un implante de mamas o una rinoplastia, inyectar bótox o ácido hialurónico, &c.) y lo que no (un aumento de la altura por encima de lo normal, capacidad de ver en la oscuridad o de respirar bajo el agua, &c.). Por consiguiente, una terapia génica que

eliminase, en la línea somática o en la línea germinal, enfermedades como la diabetes o procesos neurodegenerativos (con las privaciones que conllevan), parece admisible (de la misma manera que lo es la vacunación que erradicó mundialmente la viruela para 1980). Pero, en cambio, no parece admisible otra terapia génica que, por ejemplo, aumentase el cociente intelectual en cien puntos o nos permitiera correr más rápido que un guepardo,<sup>77</sup> características de las que carecemos por negación, no por privación.

Desde nuestra perspectiva, sin perjuicio de que hoy día sea conceder beligerancia a ciertos delirios metafísicos (lo de liquidar nuestra especie como tal es de momento una promesa mesiánica), no habría que rechazar ni aceptar en bloque los proyectos antropotécnicos transhumanistas, sino que deberíamos juzgarlos caso por caso, apelando como criterio bioético al principio de reproducción conservadora del canon de individuo humano: «el principio bioético negativo de no alterar el genoma humano en todo aquello que atente a su propia morfología [canónica]» (Bueno 2001, 85). Gran parte del conflicto bioético con los transhumanistas tiene que ver con que sus proponentes son antes biotecnólogos que médicos. La medicina es conservadora, pero la biología es evolucionista. Para los médicos, la norma fundamental es la conservación del canon de individuo humano, razón por la cual el médico buscará separar a unos hermanos siameses o seccionar un sexto dedo a un bebé, al tiempo que el biólogo se admirará de ambas mutaciones y, si es transhumanista, pensará que el sexto dedo puede hacer del neonato un pianista excepcional (Bueno 2001).

Cuando pasamos del plano de la ética, que toma como referencia al individuo, al plano de la moral, que se fija en los grupos, y al plano de la política, que se fija en las sociedades políticas, el conflicto con el transhumanismo se recrudece; porque las medidas admisibles o inadmisibles en un plano bien pueden no serlo en otro, y viceversa (Rodríguez Tauste 2022). Esto es algo en lo que muchos analistas no reparan, al no distinguir entre ética, moral y política, como sí hace el materialismo filosófico (Bueno 1996b y 2009).

---

77 Savulescu (2012, 111), dicho sea de paso, defiende el uso de drogas –el dopaje– en el deporte, como parte de las ganas de superación típicamente humanas.

Desde un punto de vista bioético, resultan condenables todas aquellas biotecnologías que ataquen el mantenimiento de la firmeza del individuo o comporten una variación inadmisiblemente del canon de individuo humano. Por su parte, desde un punto de vista biomoral o biopolítico, deben limitarse las que pongan en peligro la preservación del grupo o la eutaxia de la sociedad política, respectivamente. Pero, entonces, como señala certeramente Rodríguez Tauste (2022, 9):

Muchas mejoras que desde el punto de vista ético estarán permitidas (o incluso se promuevan, en caso de que contribuyan a la preservación de la vida de los sujetos corpóreos humanos) podrán resultar nefastas desde el punto de vista moral. Por ejemplo, si en un país europeo desarrollado introducimos mejoras que suponen un aumento significativo de la esperanza de vida, estas mejoras constituirán un deber ético pero serán inmorales en la medida en que una población demasiado envejecida, que no tiene descendencia ni constituye una fuerza productiva necesaria, es incompatible con un débil sistema de pensiones. A su vez, puede haber mejoras moralmente aceptables que supongan una violación de las normas éticas que rigen la aplicación de mejoras, tal y como ocurriría si, pongamos por caso, la supervivencia de un pueblo exige introducir mejoras peligrosas en sus soldados para poder vencer al ejército enemigo, que amenaza con impedir de una vez por todas la preservación de dicho pueblo.

Unas contradicciones entre ética y moral que no cabe resolver salomónicamente postulando la preponderancia de una sobre otra. Es la política, a través del Derecho y en función de la eutaxia de la sociedad política de referencia (EE.UU., China, Rusia, España, &c.), la que intenta articular ética y moral de manera prudente, es decir, de modo que se garantice el buen orden en el tiempo del Estado.

#### b.2) *Argumento del mejoramiento ético y moral.*

Cuando los transhumanistas exponen su imperativo de mejora, suelen colar de matute la idea de que las mejoras orgánicas también introducirán mejoras éticas y morales en la especie humana. Pero, claro, esto es mucho decir, porque ¿cuál es el gen responsable de la eticidad o de la moralidad? ¡Que lo digan! Porque el genoma no es un conjunto bien organizado, donde cada gen tenga una sola función.

Al respecto, conviene traer a colación una obra del controvertido escritor francés Michel Houellebecq, *La posibilidad de una isla* (2005).<sup>78</sup> Esta novela fue una de las primeras que abordaron concienzudamente el tema de la ingeniería genética y del transhumanismo en ciernes, aunque su autor ya había rozado el tema en *Las partículas elementales* (1998). En la novela, los personajes protagonistas (Daniel1, Esther1...), cuyas venturas y desventuras vitales se narran descarnadamente, acaban teniendo una serie de clones rectificadas genéticamente con mejoras que los convierten en autótrofos o les confieren una menor sensibilidad al dolor. Pero los neohumanos (Daniel24, Daniel25...) habitan un futuro distópico en el que llevan vidas monótonas y reiterativas en una suerte de eterno retorno proporcionado por la ciencia y la tecnología. Dado que las tentativas de descarga mnemónica a través de un soporte informático no han terminado de funcionar, los neohumanos se dedican como solución provisional a la lectura en solitario de los relatos de vida de sus predecesores, de los humanos originales (Daniel, Esther...) a partir de los cuales fueron clonados, tratando de imaginar lo que era para el hombre vivir.

Al haber sido creado genéticamente a partir de Daniel1, Daniel24 tiene, por supuesto, los mismos rasgos, la misma cara, que su antecesor; pero a los neohumanos les resulta imposible imitar esa súbita distorsión expresiva que los humanos llamaban risa, y lo mismo les ocurre con lo que llamaban alegría o amor. La lección que destila la novela de Houellebecq, siempre capaz de meter el dedo en la llaga, es que la idea de poder mejorarnos ética y moralmente por medio de la biotecnología es gratuita. O, al menos, que el precio de mejorar al ser humano es, precisamente, el dejar de ser humanos, pues la frialdad que preside la conducta de los neohumanos se encuentra en las antípodas de lo que cualquiera denominaría vida humana.

En este punto, el transhumanismo está preso de cierto reduccionismo biológico, al pensar que la modificación genética (por ejemplo, de los genes guerreros, que supuestamente controlan la agresividad) o la modificación farmacológica (mediante la inyección de oxitocina,

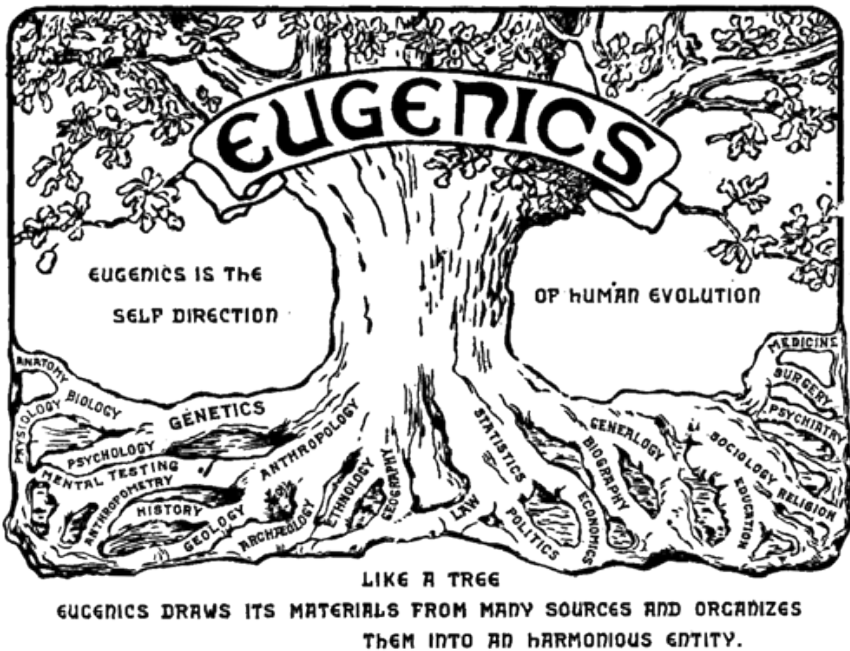
---

78 Libro, por cierto, dedicado por Houellebecq a Antonio Muñoz Ballesta, profesor de filosofía en la órbita del materialismo filosófico de Gustavo Bueno.



pongamos) pueden mejorar sustancialmente la conducta humana, soslayando el papel preponderante del ambiente y las instituciones (familia, clase social, educación, &c.). No todo está en los genes, también importan, por así decir, los memes. La antropología no se reduce a biología.

5. *Transhumanismo, fundamentalismo científico y Género humano.* Para finalizar, resaltamos que la principal crítica que puede hacerse a la doctrina transhumanista pasa por indicar que está preñada de múltiples mitos de la metafísica tradicional.



*«La eugenesia es la auto-dirección de la evolución humana.» (Lema del II Congreso Internacional de Eugenesia, celebrado en 1921, que representaba a la eugenesia como el árbol que unifica armónicamente la diversidad de disciplinas humanas, con la genética en una posición preeminente.)*

Primeramente, el movimiento transhumanista acusa el mismo fundamentalismo científico que informaba al movimiento eugenésico. De igual modo que antaño los eugenistas concebían la eugenesia como «la autodirección de la evolución humana», hogaño lo hacen los

transhumanistas con el transhumanismo. Si los eugenistas lo fiaban todo a la genética y la estadística, los transhumanistas confían en la inteligencia artificial y la biotecnología para modificar científicamente a su antojo el curso de la evolución de la especie humana.

Y, en segundo lugar, los transhumanistas tienden con frecuencia a hablar en nombre del Género humano. Pero la Humanidad, más allá de la categoría zoológica no existe, porque carece de unidad y, a lo sumo, su unidad es aureolar (se dará o no se dará en el futuro). La Humanidad, en sentido monista, no existe; porque los hombres, en cuanto desbordamos la categoría biológica, aparecemos formando pluralidades discontinuas, recortadas por lenguas y culturas diferentes, por sociedades políticas distintas. En otras palabras, como decíamos al inicio, no existe el hombre, sino que lo que existen son los españoles, los ingleses, los chinos, &c. Y, en esto, al olvidar por completo la dialéctica entre Estados e Imperios, el transhumanismo tropieza exactamente en la misma piedra que el humanismo. De hecho, al igual que O'Connor/More, Bostrom presume de que el transhumanismo es un desarrollo del humanismo secularizado y la Ilustración (Hansell & Grassie 2010, 55). Al pensar que existe el Hombre (así, con mayúscula y en singular), aunque ahora no hable de la Humanidad sino de la Humanidad+, el transhumanismo respira, en el fondo, el mismo mito que el humanismo que ambiciona superar, que es lo que queríamos demostrar.

Anexo I  
*«A Proposal for the Dartmouth Summer Research  
Project on Artificial Intelligence»*

(Una propuesta para el proyecto de investigación de verano en la Universidad de Dartmouth sobre inteligencia artificial)<sup>79</sup>

J. McCarthy, Dartmouth College  
M. L. Minsky, Harvard University  
N. Rochester, I.B.M. Corporation  
C. E. Shannon, Bell Telephone Laboratories

31 de agosto de 1955

Proponemos que durante el verano de 1956 tenga lugar en el Dartmouth College en Hanover, Nuevo Hampshire [EE.UU.], un estudio de 2 meses de duración para 10 personas. El estudio se desarrollará a partir de la conjetura de que cada aspecto del aprendizaje o cualquier otra característica de la inteligencia puede, en principio, ser descrito con tanta precisión que puede fabricarse una máquina que lo simule. Se intentará averiguar cómo fabricar máquinas que utilicen el lenguaje, formen abstracciones y conceptos, resuelvan los tipos de problemas ahora reservados para los seres humanos y mejoren por sí mismas. Creemos que puede llevarse a cabo un avance significativo en uno o más de estos problemas si un

---

79 Fuente: <http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf>

grupo de científicos cuidadosamente seleccionados trabajan en ello conjuntamente durante un verano.

Los siguientes son algunos aspectos del problema de la inteligencia artificial:

### **1. Computadoras automáticas**

Si una máquina puede hacer un trabajo, entonces se puede programar una calculadora automática para simular la máquina. Las velocidades y las capacidades de memoria de las computadoras actuales pueden ser insuficientes para simular muchas de las funciones superiores del cerebro humano, pero el principal obstáculo no es la falta de capacidad de las máquinas, sino nuestra incapacidad para escribir los programas aprovechando al máximo lo que tenemos.

### **2. ¿Cómo se puede programar una computadora para emplear un idioma?**

Puede especularse que una gran parte del pensamiento humano consiste en manipular palabras de acuerdo con las reglas del razonamiento y las reglas de la conjetura. Desde este punto de vista, formar una generalización consiste en admitir una nueva palabra y algunas reglas por las cuales las oraciones que la contienen implican y son implicadas por otras. Esta idea nunca se ha formulado con mucha precisión ni se han elaborado ejemplos.

### **3. Redes neuronales**

¿Cómo se puede disponer un conjunto de neuronas (hipotéticas) para formar conceptos? Uttley, Rashevsky y su grupo, Farley y Clark, Pitts y McCulloch, Minsky, Rochester y Holland, y otros, han realizado un trabajo teórico y experimental considerable sobre este problema. Se han obtenido resultados parciales, pero el problema necesita más trabajo teórico.

### **4. Teoría del tamaño de un cálculo**

Si se nos presenta un problema bien definido (uno para el cual es posible probar mecánicamente si una respuesta propuesta es válida o no), una forma de resolverlo es probar ordenadamente todas las respuestas posibles. Este método es ineficiente y para excluirlo se debe tener algún

criterio de eficiencia de cálculo. Algunas consideraciones mostrarán que para obtener una medida de la eficiencia de un cálculo es necesario tener a mano un método para medir la complejidad de los dispositivos de cálculo, lo que a su vez se puede hacer si uno tiene una teoría de la complejidad de las funciones. Shannon, y también McCarthy, han obtenido algunos resultados parciales sobre este problema.

### **5. Automejora**

Probablemente una máquina verdaderamente inteligente llevará a cabo actividades que bien pueden describirse como de automejora. Se han propuesto algunos esquemas para hacer esto y vale la pena estudiarlos más a fondo. Parece probable que esta cuestión también pueda estudiarse de manera abstracta.

### **6. Abstracciones**

Se pueden definir claramente varios tipos de «abstracción» y varios otros menos claramente. Un intento directo de clasificarlos y describir métodos mecánicos para formar abstracciones a partir de datos sensoriales y de otro tipo valdría la pena.

### **7. Aleatoriedad y creatividad**

Una conjetura bastante atractiva pero claramente incompleta es que la diferencia entre el pensamiento creativo y el pensamiento competente sin imaginación radica en la inyección de cierta aleatoriedad. La aleatoriedad debe estar guiada por la intuición para ser eficiente. En otras palabras, la conjetura guiada o la corazonada incluyen la aleatoriedad controlada en un pensamiento ordenado.

Además de los anteriores problemas formulados colectivamente para el estudio, hemos pedido a los participantes que describan en qué trabajarán. Se adjuntan las declaraciones de los cuatro autores del proyecto [ver más abajo].

Proponemos organizar el trabajo del grupo de la siguiente manera:

A los potenciales participantes se les enviarán copias de esta propuesta y se les preguntará si les gustaría trabajar en el problema de la inteligencia artificial en el grupo y, de ser así, en qué les gustaría trabajar. Las invitaciones serán hechas por el comité organizador sobre

la base de la estimación de la potencial contribución del individuo en cuestión al trabajo del grupo. Los integrantes harán circular sus trabajos previos y sus ideas para atacar los problemas durante los meses anteriores al período de trabajo del grupo.

Durante la reunión habrá seminarios de investigación regulares y la oportunidad para que los miembros trabajen individualmente y en pequeños grupos informales.

Los autores de esta propuesta son:

1. **C. E. Shannon**, matemático, Bell Telephone Laboratories. Shannon desarrolló la teoría estadística de la información, la aplicación del cálculo proposicional a los circuitos de conmutación y tiene resultados sobre la síntesis eficiente de circuitos de conmutación, el diseño de máquinas que aprenden, la criptografía y la teoría de las máquinas de Turing. Él y J. McCarthy están coeditando una monografía de *Annals of Mathematics* sobre «teoría de autómatas».

2. **M. L. Minsky**, Harvard Junior Fellow en Matemáticas y Neurología. Minsky ha construido una máquina para simular el aprendizaje mediante redes nerviosas y ha escrito una tesis doctoral de matemáticas en Princeton titulada «Las redes neuronales y el problema del modelo cerebral» que incluye resultados en la teoría del aprendizaje y la teoría de las redes neuronales aleatorias.

3. **N. Rochester**, gerente de investigación de la información, IBM Corporation, Poughkeepsie, Nueva York. Rochester se ocupó del desarrollo del radar durante siete años y de la maquinaria informática durante otros siete años. Él y otro ingeniero fueron conjuntamente responsables del diseño del IBM 701, que es una computadora automática a gran escala que se usa ampliamente en la actualidad. Desarrolló algunas de las técnicas de programación automática que se utilizan ampliamente en la actualidad y se ha ocupado de los problemas de cómo hacer que las máquinas realicen tareas que anteriormente solo podían realizar las personas. También ha trabajado en simulación de redes nerviosas con especial énfasis en el uso de computadoras para probar teorías en neurofisiología.

4. **J. McCarthy**, profesor de matemáticas, Dartmouth College. McCarthy ha trabajado en una serie de cuestiones relacionadas con la naturaleza matemática de los procesos de pensamiento, incluida la teoría de las máquinas de Turing, la velocidad de las computadoras, la

relación de un modelo de cerebro con su entorno y el uso de lenguajes por parte de las máquinas. Algunos resultados de este trabajo se incluyen en el monográfico de *Annals* de próxima publicación editado por Shannon y McCarthy. El otro trabajo de McCarthy ha sido en el campo de las ecuaciones diferenciales.

Se solicita a la Fundación Rockefeller que brinde apoyo financiero para el proyecto de la siguiente manera:

1. Salarios de 1200\$ para cada docente participante que no esté financiado por su propia organización. Se espera, por ejemplo, que los participantes de los Laboratorios Bell y de IBM reciban el apoyo de estas organizaciones, mientras que los de Dartmouth y Harvard requerirán el apoyo de la Fundación.

2. Salarios de 700\$ para hasta dos estudiantes graduados.

3. Billetes de tren para los participantes que vengan de lejos.

4. Alquiler para personas que estén alquilando simultáneamente en otro lugar.

5. Gastos de secretaria de 650\$, 500\$ para una secretaria y 150\$ para gastos por duplicación.

6. Gastos de organización de 200\$ (incluye los gastos de reproducción del trabajo preliminar de los participantes y los viajes necesarios para la organización).

7. Gastos para la visita de dos o tres personas por un período breve.

Gastos estimados

6 salarios de 1200	7200\$
2 salarios de 700	1400
8 gastos de viaje y de alquiler con un promedio de 300	2400
Gastos de secretaría y organización	850
Gastos de viaje adicionales	600
Contingencias	550
	<hr/>
	13.500\$

## **Propuesta de investigación de C. E. Shannon**

Me gustaría dedicar mi investigación a uno o ambos de los temas enumerados a continuación. Si bien espero hacerlo, es posible que debido a cuestiones personales no pueda asistir durante los dos meses completos. No obstante, tengo la intención de estar allí todo el tiempo que sea posible.

1. Aplicación de los conceptos de la teoría de la información a las máquinas de cómputo y los modelos cerebrales. Un problema básico en la teoría de la información es el de transmitir información de manera fiable a través de un canal ruidoso. Un problema análogo en las máquinas informáticas es el de la computación fiable utilizando elementos no fiables. Este problema ha sido estudiado por Von Neumann para la conjunción de Sheffer, y por Shannon y Moore para los relés; pero todavía restan muchas preguntas abiertas. El problema para varios elementos, el desarrollo de conceptos similares a la capacidad del canal, el análisis más preciso de los límites superior e inferior de la redundancia requerida, &c., se encuentran entre los temas importantes. Otra pregunta se relaciona con la teoría de las redes de información donde la información fluye en muchos circuitos cerrados (en contraste con el canal unidireccional simple que generalmente se considera en la teoría de la comunicación). Las cuestiones de demora se vuelven muy importantes en el caso del ciclo cerrado, y parece necesario un enfoque completamente nuevo. Esto probablemente involucraría conceptos tales como entropías parciales cuando se conoce una parte de la historia pasada de un conjunto de mensajes.
2. El modelo combinado cerebro-entorno para los autómatas. En general, una máquina o un animal solo puede adaptarse u operar en una clase limitada de entornos. Incluso el complejo cerebro humano primero se adapta a los aspectos más simples de su entorno y gradualmente desarrolla características más complejas. Propongo estudiar la síntesis de modelos cerebrales mediante el desarrollo paralelo de una serie de entornos (teóricos) combinados con los modelos cerebrales correspondientes que se adaptan a ellos. El énfasis aquí está en aclarar el modelo ambiental y representarlo como una estructura matemática. A menudo, cuando



hablamos de inteligencia mecanizada, pensamos en máquinas que realizan las actividades de pensamiento humano más avanzadas: probar teoremas, escribir música o jugar al ajedrez. Propongo aquí comenzar con lo simple, cuando el entorno no sea hostil (simplemente indiferente) ni complejo, y avanzar a través de una serie de etapas fáciles en la dirección de estas actividades avanzadas.

### **Propuesta de investigación de M. L. Minsky**

No es difícil diseñar una máquina que presente el siguiente tipo de aprendizaje. La máquina está provista de canales de entrada y salida y un medio interno para proporcionar respuestas de salida variadas a las entradas, de tal manera que la máquina puede ser *entrenada* mediante un proceso de *ensayo y error* para adquirir uno de un rango de funciones de entrada-salida. Tal máquina, cuando se coloca en un ambiente apropiado y se le da un criterio de «éxito» o «fracaso», puede ser entrenada para exhibir un comportamiento de «búsqueda de metas». A menos que la máquina cuente con, o sea capaz de desarrollar, una forma de abstraer material sensorial, puede avanzar a través de un entorno complicado solo a través de pasos dolorosamente lentos y, en general, no alcanzará un alto nivel de comportamiento.

Ahora dejemos que el criterio del éxito no sea simplemente la aparición de un patrón de actividad deseado en el canal de salida de la máquina, sino más bien la realización de una manipulación dada en un entorno dado. Entonces, en cierto modo, la situación motora parece ser dual de la situación sensorial, y el progreso puede ser razonablemente rápido solo si la máquina es igualmente capaz de ensamblar un conjunto de «abstracciones motoras» que relacionen su actividad de salida con los cambios en el entorno. Tales «abstracciones motoras» sólo pueden ser valiosas si se relacionan con cambios en el entorno que la máquina puede detectar como cambios en la situación sensorial, es decir, si están relacionadas, a través de la estructura del entorno, con las abstracciones sensoriales que la máquina está utilizando.

He estado estudiando tales sistemas durante algún tiempo y creo que si se puede diseñar una máquina en la que las abstracciones sensoriales y motoras, tal como se forman, puedan satisfacer ciertas relaciones, ello puede resultar en un alto orden de comportamiento. Estas relaciones involucran el emparejamiento de abstracciones motoras con abstracciones sensoriales de tal manera que se produzcan nuevas situaciones sensoriales que representen los cambios en el entorno que podrían esperarse si el acto motor correspondiente realmente tuviera lugar.

El resultado importante que se buscaría sería que la máquina tendiera a construir dentro de sí misma un modelo abstracto del entorno en el que se encuentra. Si se le presenta un problema, primero podría explorar soluciones dentro del modelo abstracto interno del entorno y luego intentar experimentos externos. Debido a este estudio interno preliminar, estos experimentos externos parecerían ser bastante inteligentes, y el comportamiento tendría que ser considerado como algo «imaginativo».

En mi disertación doctoral se describe una propuesta muy tentativa de cómo se podría hacer esto y tengo la intención de seguir trabajando en esta dirección. Espero que para el verano de 1956 tenga un modelo de tal máquina bastante cercano a la etapa de programación en una computadora.

### **Propuesta de investigación de N. Rochester**

#### **Originalidad en el rendimiento de una máquina**

Al escribir un programa para una calculadora automática, normalmente se proporciona a la máquina un conjunto de reglas para cubrir cada contingencia que pueda surgir y confrontar a la máquina. Uno espera que la máquina siga este conjunto de reglas servilmente y que no muestre originalidad ni sentido común. Además, uno sólo se molesta consigo mismo cuando la máquina se confunde, porque las reglas que le ha proporcionado sean ligeramente contradictorias. Finalmente, al escribir programas para máquinas, a veces uno debe abordar los problemas de una manera muy laboriosa, mientras que si la máquina tuviera un poco de intuición o pudiera hacer conjeturas

razonables, la solución del problema podría ser bastante directa. Este artículo describe una conjetura sobre cómo hacer que una máquina se comporte de una manera algo más sofisticada en el área general sugerida anteriormente. El documento analiza un problema en el que he estado trabajando esporádicamente durante unos cinco años y que deseo continuar en el Proyecto de Inteligencia Artificial el próximo verano.

### **El proceso de invención o descubrimiento**

Vivir en el ambiente de nuestra cultura nos proporciona procedimientos para resolver muchos problemas. Aún no está claro cómo funcionan estos procedimientos, pero discutiré este aspecto del problema en términos de un modelo sugerido por Craik [1]. Él sugiere que la acción mental consiste básicamente en construir pequeños motores dentro del cerebro que pueden simular y así predecir abstracciones relacionadas con el entorno.

De este modo, la solución de un problema que ya se entiende se hace de la siguiente manera:

1. El entorno proporciona datos a partir de los cuales se forman ciertas abstracciones.
2. Las abstracciones junto con ciertos hábitos o impulsos internos proporcionan:
  - 2.1. Una definición de un problema en términos de la condición deseada a lograr en el futuro, una meta.
  - 2.2. Una acción sugerida para resolver el problema.
  - 2.3. Estimulación para despertar en el cerebro el motor que corresponde a esta situación.
3. Luego, el motor funciona para predecir a qué conducirá esta situación ambiental y la reacción propuesta.
4. Si la predicción corresponde a la meta, el individuo procede a actuar como se indica.

La predicción corresponderá a la meta si vivir en el ambiente de su cultura le ha proporcionado al individuo la solución al problema. En cuanto al individuo como una calculadora de programa almacenado, el programa contiene reglas para cubrir esta contingencia particular.

Para una situación más compleja, las reglas pueden ser más complicadas. Las reglas pueden requerir probar cada una de un conjunto de acciones posibles para determinar cuál proporcionó la solución. Un conjunto de reglas aún más complejo podría proporcionar incertidumbre sobre el entorno, como por ejemplo, al jugar al tres en raya, uno no solo debe considerar su próximo movimiento, sino también los diversos movimientos posibles del entorno (su oponente).

Ahora considérese un problema para el cual ningún individuo en la cultura tiene una solución y que ha resistido los esfuerzos de solución. Este podría ser un típico problema científico actual sin resolver. El individuo podría tratar de resolverlo y encontrar que cada acción razonable condujo al fracaso. En otras palabras, el programa almacenado contiene reglas para la solución de este problema, pero las reglas son ligeramente incorrectas.

Para resolver este problema, el individuo tendrá que hacer algo que es irrazonable o inesperado a juzgar por la herencia de sabiduría acumulada por la cultura. Podría obtener tal comportamiento probando diferentes cosas al azar, pero tal enfoque generalmente sería demasiado ineficiente. Por lo general, hay demasiados cursos de acción posibles de los cuales solo una pequeña fracción es aceptable. El individuo necesita una corazonada, algo inesperado pero no del todo razonable. Algunos problemas, a menudo los que son bastante nuevos y no han resistido mucho esfuerzo, necesitan solo un poco de aleatoriedad. Otros, a menudo aquellos que se han resistido durante mucho tiempo a la solución, necesitan una desviación realmente extraña de los métodos tradicionales. Un problema cuya solución requiere originalidad podría dar paso a un método de solución que implicara aleatoriedad.

En términos del modelo S de Craik [1], el motor que debería simular el entorno al principio no simula correctamente. Por lo tanto, es necesario probar varias modificaciones del motor hasta encontrar una que haga hacer lo que se necesita.

En lugar de describir el problema en términos de un individuo en su cultura, podría haberse descrito en términos del aprendizaje de un individuo inmaduro. Cuando al individuo se le presenta un problema fuera del alcance de su experiencia, debe superarlo de manera similar.

Hasta ahora, el enfoque práctico más cercano que usa este método en la solución de problemas con máquinas es una extensión del

método de Montecarlo. En el problema habitual que es apropiado para Montecarlo hay una situación que se malinterpreta gravemente y que tiene demasiados factores posibles y uno es incapaz de decidir qué factores ignorar al elaborar una solución analítica. Así, el matemático hace que la máquina haga unos cuantos miles de experimentos aleatorios. Los resultados de estos experimentos proporcionan una suposición aproximada de cuál puede ser la respuesta. La extensión del método de Montecarlo es usar estos resultados como una guía para determinar qué ignorar para simplificar el problema lo suficiente como para obtener una solución analítica aproximada.

Cabría preguntarse por qué el método debería incluir la aleatoriedad. ¿Por qué el método no debería ser probar cada posibilidad en el orden de la probabilidad de que el estado actual del conocimiento prediga su éxito? Para el científico rodeado por el entorno proporcionado por su cultura, es posible que un solo científico sea poco probable que resuelva el problema en su vida, por lo que se necesitan los esfuerzos de muchos. Si usan la aleatoriedad, podrían trabajar todos a la vez sin una duplicación completa del esfuerzo. Si usaran el sistema, requerirían una comunicación increíblemente detallada. Para el individuo que madura en competencia con otros individuos, los requisitos de la estrategia mixta (utilizando la terminología de la teoría de juegos) favorecen la aleatoriedad. Para la máquina, probablemente se necesitará aleatoriedad para superar la miopía y los prejuicios del programador. Aunque claramente no se ha probado la necesidad de la aleatoriedad, hay muchas pruebas a su favor.

### **La máquina con aleatoriedad**

Para escribir un programa para hacer que una calculadora automática use la originalidad, no servirá introducir aleatoriedad sin usar la previsión. Si, por ejemplo, uno escribiera un programa para que una vez cada 10.000 pasos la calculadora generara un número aleatorio y lo ejecutara como una instrucción, el resultado probablemente sería un caos. Luego, después de cierta cantidad de caos, la máquina probablemente intentaría algo prohibido o ejecutaría una instrucción de parada y el experimento terminaría.

Sin embargo, dos enfoques parecen ser razonables. Uno de ellas es encontrar cómo el cerebro logra hacer este tipo de cosas y copiarlas.

El otro es tomar alguna clase de problemas reales que requieran originalidad en su solución e intentar encontrar la manera de escribir un programa para resolverlos en una calculadora automática. Cualquiera de estos enfoques es probable que finalmente tenga éxito. Sin embargo, no está claro cuál sería más rápido ni cuántos años o generaciones tomaría cada uno. Hasta ahora, la mayor parte de mi esfuerzo en este sentido se ha centrado en el primer enfoque, porque sentí que sería mejor dominar todo el conocimiento científico relevante para poder trabajar en un problema tan difícil, y ya era bastante consciente del estado actual de las calculadoras y del arte de programarlas.

El mecanismo de control del cerebro es claramente muy diferente del mecanismo de control de las calculadoras actuales. Un síntoma de la diferencia es la forma de fallar. Un fallo en una calculadora característicamente produce algo bastante irrazonable. Un error en la memoria o en la transmisión de datos es tan probable que sea en el dígito más significativo como en el menor. Un error en el control puede hacer casi cualquier cosa. Podría ejecutar la instrucción incorrecta u operar una unidad de entrada-salida incorrecta. Por otro lado, los errores humanos en el habla tienden a dar como resultado declaraciones que casi tienen sentido (considere a alguien que está casi dormido, ligeramente borracho o ligeramente febril). Quizás el mecanismo del cerebro es tal que un ligero error en el razonamiento introduce la aleatoriedad de la manera adecuada. Quizás el mecanismo que controla el orden serial en el comportamiento [2] guía el factor aleatorio para mejorar la eficiencia de los procesos imaginativos sobre la pura aleatoriedad.

Se ha trabajado en la simulación de redes neuronales en nuestra calculadora automática. Uno de los propósitos era ver si sería posible introducir la aleatoriedad de una manera apropiada. Parece haber resultado que hay demasiados vínculos desconocidos entre la actividad de las neuronas y la resolución de problemas para que este enfoque funcione todavía. Los resultados han arrojado algo de luz sobre el comportamiento de las redes y las neuronas, pero no han arrojado una forma de resolver problemas que requieran originalidad.

Un aspecto importante de este trabajo ha sido un esfuerzo por hacer que la máquina forme y manipule conceptos, abstracciones, generalizaciones y nombres. Se hizo un intento de probar una teoría

[3] de cómo lo hace el cerebro. La primera serie de experimentos ocasionó una revisión de ciertos detalles de la teoría. El segundo conjunto de experimentos está en progreso. Para el próximo verano este trabajo estará terminado y se habrá escrito un informe final.

Mi plan es tratar de escribir un programa para resolver problemas que son miembros de una clase limitada de problemas que requieren originalidad en su solución. Es demasiado pronto para predecir en qué etapa estaré el próximo verano o, simplemente, cómo definiré entonces el problema inmediato. Sin embargo, el problema subyacente que se describe en este documento es lo que pretendo abordar. En una sola frase, el problema es: ¿cómo puedo hacer una máquina que muestre originalidad en la solución de problemas?

### Referencias

1. K. J. W. Craik, *The Nature of Explanation*, Cambridge University Press, 1943 (reimpreso en 1952), p. 92.
2. K. S. Lashley, «The Problem of Serial Order in Behavior», en *Cerebral Mechanism in Behavior, the Hixon Symposium*, editado por L. A. Jeffress, John Wiley & Sons, Nueva York, pp. 112-146, 1951.
3. D. O. Hebb, *The Organization of Behavior*, John Wiley & Sons, Nueva York, 1949.

### **Propuesta de investigación de John McCarthy**

Durante el próximo año y durante el Proyecto de Investigación de Verano sobre Inteligencia Artificial, propongo estudiar la relación del lenguaje con la inteligencia. Parece claro que la aplicación directa de métodos de prueba y error a la relación entre los datos sensoriales y la actividad motora no conducirá a ningún comportamiento muy complicado. Más bien, es necesario que los métodos de prueba y error se apliquen a un mayor nivel de abstracción. La mente humana aparentemente usa el lenguaje como medio para manejar fenómenos complicados. Los procesos de prueba y error en un nivel superior frecuentemente toman la forma de formular conjeturas y probarlas. El idioma inglés tiene una serie de propiedades de las que carecen todos los idiomas formales descritos hasta ahora.

1. Los argumentos en inglés complementados con matemáticas informales pueden ser concisos.

2. El inglés es universal en el sentido de que puede configurar cualquier otro idioma dentro del inglés y luego usar ese idioma donde sea apropiado.

3. El usuario del inglés puede referirse a sí mismo en él y formular afirmaciones sobre su progreso en la solución del problema en el que está trabajando.

4. Además de las reglas de prueba, el inglés, si estuviera completamente formulado, tendría reglas de conjetura.

Los lenguajes lógicos formulados hasta ahora han sido listas de instrucciones para hacer que las computadoras realicen cálculos especificados de antemano o bien para la formalización de partes de las matemáticas. Estos últimos han sido construidos de manera que:

1. sean fácilmente descritos en matemáticas informales,
2. permitan la traducción de proposiciones matemáticas informales al lenguaje,
3. faciliten la discusión sobre las demostraciones de (???)

No se ha hecho ningún intento de hacer demostraciones en lenguajes artificiales tan breves como las demostraciones informales. Por lo tanto, parece deseable intentar construir un lenguaje artificial que una computadora pueda programar para usar en problemas que requieren conjeturas y autorreferencia. Debe corresponder al inglés en el sentido de que las frases cortas en inglés sobre el tema dado deben tener correspondencias cortas en el lenguaje y también los argumentos cortos o los argumentos conjeturales. Espero tratar de formular un lenguaje que tenga estas propiedades y además contenga las nociones de objeto físico, suceso, &c., con la esperanza de que usando este lenguaje sea posible programar una máquina para aprender a jugar bien a diversos juegos y a hacer otras tareas.<sup>80</sup>

---

80 En 1959, McCarthy desarrolló el lenguaje de programación LISP, justo un año después de que apareciera FORTRAN, el sistema de traducción de fórmulas matemáticas de IBM (*N. del T.*).



## **Personas interesadas en el problema de la inteligencia artificial**

El propósito de esta lista es que los que están en ella sepan quién está interesado en recibir documentos sobre el problema. Las personas primeramente apuntadas recibirán copias del informe del Proyecto de Verano de Dartmouth sobre Inteligencia Artificial.

La lista consta de personas que participarán o visitarán el Proyecto de Investigación de Verano de Dartmouth sobre Inteligencia Artificial, o que se sabe que están interesadas en el tema [no todas las personas enumeradas asistieron a la conferencia de Dartmouth]. Se está enviando a la gente apuntada primeramente y a algunos otros.

Para el presente propósito, el problema de la inteligencia artificial se toma como el de hacer que una máquina se comporte de maneras que se llamarían inteligentes si un humano se comportara así.

Próximamente se publicará una lista revisada, por lo que cualquier otra persona interesada en incluirse en la lista o cualquier persona que desee cambiar su dirección debe escribir a:

John McCarthy, Departamento de Matemáticas, Dartmouth College, Hanover, NH.

La lista consta de: [aparte de los nombres de los organizadores anteriormente mencionados, aparecen Allen Newell, Warren McCulloch, John Nash, Abraham Robinson, Norman Shapiro, John G. Kemeny o Herbert Simon, entre otros interesados en EE.UU., Inglaterra y Canadá].



## Anexo II

### «Pause Giant AI Experiments: An Open Letter»

(Pausar los experimentos gigantes de IA: una carta abierta)<sup>81</sup>

22 de marzo de 2023

Hacemos un llamamiento a todos los laboratorios de IA para que pausen inmediatamente durante al menos 6 meses el entrenamiento de los sistemas de IA más potentes que GPT-4.

Los sistemas de IA con inteligencia humana competente pueden plantear riesgos profundos para la sociedad y la Humanidad, como lo demuestra una extensa investigación y lo reconocen los principales laboratorios de IA.<sup>82</sup> Como se establece en los Principios de IA de Asilomar,<sup>83</sup> ampliamente respaldados, la IA avanzada podría representar un cambio profundo en la historia de la vida en la Tierra y debe planificarse y administrarse con el cuidado y los recursos correspondientes. Desafortunadamente, este nivel de planificación y gestión no está ocurriendo, dado que en los últimos meses los laboratorios de IA han entrado en una carrera fuera de control por desarrollar e implementar mentes digitales cada vez más poderosas, que nadie, ni siquiera sus creadores, pueden entender, predecir o controlar de forma fiable.

---

81 Fuente: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

82 La carta refiere a los trabajos al respecto de Nick Bostrom, Stuart Russell, Max Tegmark y otros autores, así como a las opiniones de los CEOs de OpenAI y DeepMind (*N. del T.*).

83 Cf. Anexo III (*N. del T.*).

Los sistemas de IA contemporáneos están volviéndose competentes como los humanos para realizar tareas generales,<sup>84</sup> y debemos preguntarnos: ¿Deberíamos dejar que las máquinas inunden nuestros canales de información con propaganda y falsedad? ¿Deberíamos automatizar todos los trabajos, incluidos los gratificantes? ¿Deberíamos desarrollar mentes no humanas que eventualmente podrían superarnos en número, ser más inteligentes y reemplazarnos? ¿Deberíamos arriesgarnos a perder el control de nuestra civilización? Tales decisiones no deben delegarse en líderes tecnológicos no elegidos. Los sistemas potentes de IA deben desarrollarse solo una vez que estemos seguros de que sus efectos serán positivos y los riesgos asumibles. Esta confianza debe estar bien justificada y aumentar con la magnitud de los potenciales efectos del sistema. La reciente declaración de OpenAI con respecto a la inteligencia artificial general establece que «en algún momento, puede ser importante obtener una revisión independiente antes de comenzar a entrenar sistemas futuros, y para los esfuerzos más avanzados acordar limitar la tasa de crecimiento de la computación utilizada para crear nuevos modelos». Estamos de acuerdo. Ese momento es ahora.

Por lo tanto, hacemos un llamamiento a todos los laboratorios de IA para que pausen de inmediato durante al menos 6 meses el entrenamiento de los sistemas de IA más potentes que GPT-4. Esta pausa debe ser pública y verificable, e incluir a todos los actores clave. Si tal pausa no se puede promulgar rápidamente, los gobiernos deberían intervenir e instituir una moratoria.

Los laboratorios de IA y expertos independientes deberían aprovechar esta pausa para desarrollar e implementar conjuntamente una serie de protocolos de seguridad compartidos para el diseño y el desarrollo avanzados de la IA, que sean rigurosamente auditados y supervisados por expertos externos independientes. Estos protocolos deben garantizar que los sistemas que se adhieren a ellos sean seguros más allá de toda duda razonable.<sup>85</sup> Esto *no* significa una pausa en el desarrollo de la IA en general, simplemente un paso atrás en la

---

84 La carta remite a Bubeck & al. (2023) (*N. del T.*).

85 Existe un amplio precedente legal; por ejemplo, los Principios de IA de la OCDE, ampliamente adoptados, exigen que los sistemas de IA «funcionen adecuadamente y no planteen un riesgo de seguridad irrazonable».

peligrosa carrera hacia impredecibles modelos de caja negra cada vez más grandes y con capacidades emergentes.

La investigación y el desarrollo de la IA deben reenfocarse en hacer que los sistemas potentes y de última generación de hoy en día sean más precisos, seguros, interpretables, transparentes, robustos, alineados, confiables y leales.

Paralelamente, los desarrolladores de IA deben trabajar con los legisladores para acelerar drásticamente el desarrollo de sistemas robustos de gobierno de la IA. Estos deben incluir como mínimo: nuevas y capaces autoridades reguladoras dedicadas a la IA; la supervisión y el seguimiento de los sistemas de IA con altas capacidades y gran poder computacional; sistemas de procedencia y de marcas de agua que ayuden a distinguir lo real de lo artificial y a rastrear los modelos implicados; un sólido ecosistema de auditoría y certificación; responsabilidad por los daños causados por la IA; financiación pública sólida para la investigación técnica en seguridad de la IA; e instituciones bien dotadas para hacer frente a las dramáticas perturbaciones económicas y políticas (especialmente para la democracia) que provocará la IA.

La Humanidad puede disfrutar de un futuro floreciente con la IA. Habiendo tenido éxito en la creación de poderosos sistemas de IA, ahora podemos disfrutar de un «verano de la IA» en el que cosechemos las recompensas, diseñemos estos sistemas para el claro beneficio de todos y le demos a la sociedad la oportunidad de adaptarse. La sociedad ha hecho una pausa en otras tecnologías con efectos potencialmente catastróficos en la sociedad.<sup>86</sup> Podemos hacerlo aquí. Disfrutemos de un largo verano de la IA, no nos apresuremos a caer sin estar preparados.

---

<sup>86</sup> Los ejemplos incluyen la clonación humana, la modificación de la línea germinal humana, la investigación médica que altera genéticamente el organismo y la eugenesia.



## Anexo III «Asilomar AI Principles»

(Los 23 Principios de la Inteligencia Artificial coordinados por el Instituto para el Futuro de la Vida y desarrollados en la conferencia celebrada en Asilomar en 2017)<sup>87</sup>

La inteligencia artificial ya ha proporcionado herramientas beneficiosas que personas de todo el mundo utilizan todos los días. Su desarrollo continuo, guiado por los siguientes principios, ofrecerá increíbles oportunidades para ayudar y empoderar a las personas en las décadas y siglos venideros.

### **Cuestiones de investigación**

1. Meta de la investigación: el objetivo de la investigación en IA no debería ser crear inteligencia sin dirigir, sino inteligencia beneficiosa.

2. Financiación de la investigación: la inversión en IA debería ir acompañada de fondos para investigar en asegurar su uso beneficioso, incluyendo cuestiones espinosas sobre ciencias de la computación, economía, derecho, ética y estudios sociales, tales como:

- ¿Cómo podemos hacer que los futuros sistemas de IA sean muy robustos, de modo que hagan lo que queremos sin funcionar mal ni ser pirateados?

---

87 Fuente: <https://futureoflife.org/ai-principles/>

- ¿Cómo podemos aumentar nuestra prosperidad a través de la automatización y al mismo tiempo mantener los recursos y el propósito de las personas?

- ¿Cómo podemos actualizar nuestros sistemas legales para que sean más justos y eficientes, para seguir el ritmo de la IA y para gestionar los riesgos asociados con ella?

- ¿Con qué conjunto de valores debería alinearse la IA y qué estatus legal y ético debería tener?

3. Enlace entre ciencia y política: debería haber un intercambio constructivo y sano entre los investigadores en IA y los políticos.

4. Cultura de la investigación: una cultura de cooperación, confianza y transparencia debería ser fomentada entre los investigadores y desarrolladores de IA.

5. Evitar las carreras: los equipos que estén desarrollando sistemas de IA deberían cooperar activamente para evitar chapuzas en los estándares de seguridad.

## **Ética y valores**

6. Seguridad: los sistemas de IA deberían ser seguros a lo largo de su vida operativa, y verificables donde sea aplicable y posible.

7. Transparencia en los fallos: si un sistema de IA causa daño, debería ser posible determinar por qué.

8. Transparencia judicial: cualquier intervención de un sistema autónomo en una decisión debería ir acompañada de una explicación satisfactoria y auditable por parte de una autoridad humana competente.

9. Responsabilidad: los diseñadores y desarrolladores de sistemas avanzados de IA son depositarios de las implicaciones morales de su uso, mal uso y acciones, con la responsabilidad y oportunidad de dar forma a dichas implicaciones.

10. Alineación de valores: los sistemas de IA altamente autónomos deberían ser diseñados para que sus metas y comportamientos puedan alinearse con los valores humanos a lo largo de sus operaciones.

11. Valores humanos: los sistemas de IA deberían ser diseñados y operados para que sean compatibles con los ideales de dignidad humana, derechos, libertades y diversidad cultural.



12. Privacidad personal: la gente debería tener el derecho de acceder, gestionar y controlar los datos que generan, dando a los sistemas de IA el poder de analizar y utilizar esa información.

13. Libertad y privacidad: la aplicación de la IA a los datos personales no puede restringir de modo poco razonable la libertad, real o sentida, de las personas.

14. Beneficio compartido: las tecnologías de IA deberían beneficiar y empoderar a tanta gente como sea posible.

15. Prosperidad compartida: la prosperidad económica creada por la IA debería ser compartida ampliamente, para el beneficio de toda la Humanidad.

16. Control humano: los seres humanos deberían escoger cómo y si delegan decisiones a los sistemas de IA, para completar objetivos escogidos por seres humanos.

17. Sin subversión: el poder conferido por el control de sistemas de IA altamente avanzados debería respetar y mejorar, más que subvertir, los procesos sociales y cívicos de los que depende la salud de la sociedad.

18. Carrera armamentística: debería ser evitada cualquier carrera armamentística de armas autónomas letales.

### **Cuestiones a largo plazo**

19. Capacidad de precaución: al no haber consenso, deberíamos evitar las asunciones contundentes sobre los límites superiores de las futuras capacidades de la IA.

20. Importancia: la IA avanzada podría representar un profundo cambio en la historia de la vida en la Tierra, y debería ser planificada y gestionada con el cuidado y los recursos adecuados.

21. Riesgos: los riesgos asociados a los sistemas de IA, especialmente los catastróficos o existenciales, deben estar sujetos a planificación y esfuerzos de mitigación equiparables a su impacto esperado.

22. Automejora recursiva: los sistemas de IA diseñados para automejorarse recursivamente o autorreplicarse de una forma que pudiera llevar al rápido incremento en su calidad o cantidad deben estar sujetos a unas estrictas medidas de control y seguridad.

23. Bien común: la superinteligencia debería ser desarrollada sólo en servicio de unos ideales éticos ampliamente compartidos y para beneficio de toda la Humanidad, más que para un Estado u organización.

## Anexo IV «*Transhumanist Declaration*»

(Declaración Transhumanista adoptada por la organización Humanity+, anteriormente conocida como la Asociación Transhumanista Mundial, en marzo de 2009)<sup>88</sup>

1. La humanidad se verá profundamente afectada por la ciencia y la tecnología en el futuro. Prevedemos la posibilidad de ampliar el potencial humano superando el envejecimiento, las limitaciones cognitivas, el sufrimiento involuntario y nuestro confinamiento en el planeta Tierra.

2. Creemos que el potencial de la humanidad aún no se ha desarrollado plenamente. Hay escenarios posibles que conducen a condiciones humanas maravillosas y que merecen la pena.

3. Reconocemos que la humanidad se enfrenta a riesgos serios, especialmente por el mal uso de las nuevas tecnologías. Hay escenarios realistas que podrían llevarnos a la pérdida de casi todo o, incluso, de todo lo que consideramos valioso. Algunos de estos escenarios son drásticos, otros son sutiles. Aunque todo progreso es cambio, no todo cambio representa progreso.

4. Es necesario invertir esfuerzo en la investigación para comprender estas perspectivas. Necesitamos deliberar cuidadosamente sobre el mejor modo de reducir los riesgos y acelerar las aplicaciones beneficiosas. También se requieren foros donde se pueda discutir

---

88 Fuente: <http://humanityplus.org/philosophy/transhumanist-declaration/>

constructivamente lo que debe hacerse, y un orden social donde se puedan tomar decisiones responsables.

5. La reducción de los riesgos para nuestra existencia como humanos y el desarrollo de medios para la preservación de la vida y la salud, el alivio del sufrimiento grave y el incremento de la previsión y la sabiduría humanas deben perseguirse como prioridades urgentes y recibir una financiación importante.

6. El diseño de políticas debe guiarse por una visión moral responsable e inclusiva, tomando en consideración tanto las oportunidades como los riesgos, respetando la autonomía y los derechos individuales, y mostrando solidaridad y preocupación por los intereses y la dignidad de todas las personas alrededor del globo. También debemos tomar en cuenta nuestra responsabilidad moral para con las generaciones venideras.

7. Abogamos por el bienestar de todo ser sensible, lo que incluye a los humanos, los animales no humanos y cualesquiera intelectos artificiales futuros, formas de vida modificadas u otras inteligencias a las que pueda dar lugar el avance tecnológico y científico.

8. Estamos a favor de que los individuos sean los que decidan y elijan sobre cómo vivir sus vidas. Esto incluye el uso de técnicas que puedan incrementar la memoria, la concentración y la energía mental; terapias para alargar la vida; tecnologías de selección reproductiva; procedimientos criogénicos; y muchas otras posibles tecnologías de modificación y mejora humana.

## Referencias bibliográficas

- Agüera y Arcas, Blaise (2021): «Do large language models understand us?», *Medium*, 16 de diciembre.
- Agüera y Arcas, Blaise & Norvig, Peter (2023): «Artificial General Intelligence Is Already Here», *Noema*, 10 de octubre.
- Alvargonzález, David (2010): «La transdisciplinariedad como mito milenarista», *Encuentros multidisciplinares*, 34, 70-79.
- Anderson, Chris (2008): «The End of Theory: The Data Deluge Makes the Scientific Method Obsolete», *Wired*, 23 de junio.
- Arkoudas, Konstantine (2023): «ChatGPT is no stochastic parrot. But it also claims that  $1 > 1$ », *Medium*, 15 de enero.
- Aschwanden, Christie (2020): «Artificial Intelligence Makes Bad Medicine Even Worse», *Wired*, 10 de enero.
- Baños Pino, Carmen (2016): *El problema de Molyneux desde el materialismo filosófico*, Pentalfa, Oviedo.
- Belda, Ignasi (2011): *Mentes, máquinas y matemáticas. La inteligencia artificial y sus retos*, RBA, España.
- Benjamins, Richard & Salazar, Idoia (2020): *El mito del algoritmo. Cuentos y cuentas de la inteligencia artificial*, Anaya, Madrid.
- Borji, Ali (2023): «A Categorical Archive of ChatGPT Failures», *Arxiv*, 3 de abril.
- Bostrom, Nick (2016 [2014]): *Superinteligencia: caminos, peligros, estrategias*, Teell, España.
- Bubeck, Sébastien & al. (2023): «Sparks of Artificial General Intelligence: Early experiments with GPT-4», *Arxiv*, 22 de marzo.
- Bueno, Gustavo (1972): *Ensayos materialistas*, Taurus, Madrid.
- (1978): «En torno al concepto de ‘ciencias humanas’. La distinción entre metodologías  $\alpha$ -operatorias y  $\beta$ -operatorias», *El Basilisco*, 2, 12-46.
- (1982): «Psicoanalistas y epicúreos», *El Basilisco*, 13, 12-39.

- (1985a): «Lenguaje y pensamiento en Platón», *Taula*, 3, 39-59.
- (1985b): «Para un análisis gnoseológico de la psicología cognitiva», *Estudios de Psicología*, 22, 103-113.
- (1989): *Cuestiones cuodlibetales sobre Dios y la religión*, Mondadori, Madrid.
- (1990): *Materia*, Pentalfa, Oviedo (versión española del artículo en la *Europäische Enzyklopädie zu Philosophie und Wissenschaften*).
- (1992): *Teoría del cierre categorial*, 5 volúmenes, Pentalfa, Oviedo.
- (1993): «En torno a la doctrina filosófica de la causalidad», en *Revista Meta*, Editorial Complutense, Madrid, 207-227.
- (1995a): *¿Qué es la ciencia?*, Pentalfa, Oviedo.
- (1995b): *¿Qué es la filosofía?*, Pentalfa, Oviedo.
- (1996a): «Información y causalidad», en *El concepto de información en las ciencias naturales y sociales*, UCM, Madrid, 11-30.
- (1996b): *El sentido de la vida. Seis lecturas de filosofía moral*, Pentalfa, Oviedo.
- (2000): «Notas sobre la técnica», *Ábaco*, 27-28, 193-206.
- (2001): *¿Qué es la Bioética?*, Biblioteca Filosofía en Español, Oviedo.
- (2004): *La vuelta a la caverna*, Ediciones B, Barcelona.
- (2005): «Ensayo de una teoría antropológica de las instituciones», *El Basilisco*, 37, 3-52.
- (2006): «Sobre los derechos de los simios», en *Zapatero y el Pensamiento Alicia*, Temas de Hoy, Barcelona, 109-158.
- (2008): «¡Dios salve la Razón!», en Benedicto XVI, Gustavo Bueno & al., *Dios salve la Razón*, Encuentro, Madrid, 57-92.
- (2009): «Ética, Moral, Bioética, Derecho», *El Catoblepas*, 93, 2.
- (2010): «Fundamentalismo científico y bioética», *El Catoblepas*, 97, 2.
- (2011): «Paz, Democracia y Razón», *El Catoblepas*, 116, 2.
- (2015): «El humanismo como ideal supremo», *El Catoblepas*, 158, 2.
- (2016 [1996]): *El mito de la Cultura*, 10ª edición, Pentalfa, Oviedo.
- Bunge, Mario (1977): «Towards a Technoethics», *The Monist*, 60/1, 96-107.
- Carabantes López, Manuel (2016): *Inteligencia artificial. Una perspectiva filosófica*, Guillermo Escolar, Madrid.
- Chakravorti, Bhaskar (2022): «Why AI Failed to Live Up to Its Potential During the Pandemic», *Harvard Business Review*, 17 de marzo.
- Chan, Stephen & Siegel, Eliot L. (2019): «Will machine learning end the viability of radiology as a thriving medical specialty?», *Br. J. Radiol.*, 92(1094), 20180416.
- Cobo Cano, Miriam & Lloret Iglesias, Lara (2023): *Inteligencia artificial y medicina*, CSIC, Madrid.
- Coeckelbergh, Mark (2021 [2020]): *Ética de la Inteligencia Artificial*, Cátedra, Madrid.
- (2023 [2022]): *La filosofía política de la inteligencia artificial*, Cátedra, Madrid.
- Comisión Europea (2018): «IA para Europa», 25 de abril.
- (2020): *Libro Blanco sobre la IA*, 19 de febrero.

- (2021): «Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts», 21 de abril.
- Crawford, Kate (2021): *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale UP, New Haven.
- D. María, Álvaro (2024): *La filosofía de Bitcoin. La caída del Estado*, Deusto, Barcelona.
- Davis, Ernest (2023): «Machine Translate fails on simple sentences», página web actualizada el 25 de marzo.
- De Lorenzo, Javier (2020): *Un mundo de artefactos*, Trotta, Madrid.
- Degli-Esposti, Sara (2023): *La ética de la inteligencia artificial*, CSIC, Madrid.
- Diéguez, Antonio (2017): *Transhumanismo*, Herder, Barcelona.
- (2021): *Cuerpos inadecuados*, Herder, Barcelona.
- Domingos, Pedro (2012): «A few useful things to know about machine learning», *Commun. ACM*, 55, 78-87.
- (2015): *The Master Algorithm*, Basic Books, Nueva York.
- Dreyfus, Hubert (1965): «Alchemy and Artificial Intelligence», informe para la corporación RAND (Santa Mónica, California).
- (1972): *What Computers Can't Do*, MIT Press, Massachusetts.
- (1992): *What Computers Still Can't Do: A Critique of Artificial Reason*, MIT Press, Massachusetts.
- Dreyfus, Hubert & Dreyfus, Stuart (1987): *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*, Free Press, Nueva York.
- EASA (2023): «EASA Artificial Intelligence Roadmap 2.0», European Union Aviation Safety Agency, Colonia, mayo.
- Echeverría, Javier (2010): «Tecnociencia, tecnoética y tecnoaxiología», *Revista Colombiana de Bioética*, 5/1, 142-152.
- Eloundou, Tyna & al. (2023): «GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models», informe para OpenAI, 27 de marzo.
- Estany, Anna & Casacuberta, David (eds.) (2021): *Epistemología e innovación en medicina*, Plaza y Valdés, Madrid.
- Floridi, Luciano (2013): «Distributed morality in an information society», *Science and Engineering Ethics*, 19(3), 727-743.
- (2021): «The European Legislation on AI: a brief Analysis of its Philosophical Approach», *Philosophy & Technology*, 34, 215-222.
- (2022): «Ultraintelligent Machines, Singularity, and Other Sci-fi Distractions about AI», *Lavoro Diritti Europa*, 3/2022, 8 de noviembre.
- (2023a): «AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models», *Philosophy & Technology*, 36, 15.
- (2023b): *The Ethics of Artificial Intelligence*, Oxford UP, Oxford.
- Floridi, Luciano & Chiriatti, M. (2020): «GPT-3: Its Nature, Scope, Limits, and Consequences», *Minds and Machines*, 30, 681-694.
- Floridi, Luciano & Sanders, J. W. (2004): «On the Morality of Artificial Agents», *Minds and Machines*, 14, 349-379.

- Fukuyama, Francis (2003 [2002]): *El fin del hombre. Consecuencias de la revolución biotecnológica*, Ediciones B, Barcelona.
- (2004): «Transhumanism», *Foreign Policy*, 144, 43-43.
- Gartner (2022): «What's New in Artificial Intelligence from the 2022 Gartner Hype Cycle», informe de la consultora, 15 de septiembre.
- Génova Fuster, Gonzalo (2023): «Inteligencia artificial: explicabilidad, racionalidad, y responsabilidad profesional del ingeniero», intervención en los 28 Encuentros de Filosofía *Filosofía, Inteligencia Artificial y Nuevas Tecnologías* de la Fundación Gustavo Bueno, 29 de septiembre.
- Génova, Gonzalo & Quintanilla Navarro, Ignacio (2018): «Discovering the principle of finality in computational machines», *Foundations of Science*, 23(4), 779-794.
- Gracia, Diego (2007): *Fundamentos de Bioética*, 3ª edición, Triacastela, Madrid.
- Grupo de Expertos de Alto Nivel sobre IA de la Unión Europea (2019a): «Ethics Guidelines for Trustworthy AI», 8 de abril.
- (2019b): «Una definición de la inteligencia artificial: Principales capacidades y disciplinas científicas», 8 de abril.
- (2020): «The Assessment List for Trustworthy AI (ALTAI)», 17 de julio.
- Grupo Europeo de Ética de la Ciencia y de las Nuevas Tecnologías (2018): «Statement on AI, Robotics, and 'Autonomous' Systems», 9 de marzo.
- Habermas, Jürgen (2002 [2001]): *El futuro de la naturaleza humana. ¿Hacia una eugenesia liberal?*, Paidós, Barcelona.
- Hansell, Gregory R. & Grassie, William (eds.) (2010): *H± Transhumanism and Its Critics, Metanexus, Estados Unidos de Norteamérica* (contiene contribuciones de Nick Bostrom, Max More, Natasha Vita-More, Don Ihde y Andrew Pickering, entre otros).
- Herrera Triguero, Francisco (2019): *Inteligencia computacional: sistemas inteligentes inspirados en la naturaleza (discurso del académico electo)*, Real Academia de Ingeniería, España.
- Jobin, Anna & al. (2019): «The global landscape of AI ethics guidelines», *Nature Machine Intelligence*, 1, 389-399.
- Julia, Luc (2019): *L'intelligence artificielle n'existe pas*, Éditions First, París.
- Kaplan, Jerry (2017 [2016]): *Inteligencia artificial. Lo que todo el mundo debe saber*, Teell, España.
- Kurzweil, Raymond (2012 [2005]): *La singularidad está cerca*, Lola Books, España.
- Lahoz Beltra, Rafael (2012): *Turing. Pensando en máquinas que piensan*, National Geographic, Rodesa.
- Lanier, Jaron (2023): «There is no A.I.», *The New Yorker*, 20 de abril.
- Lara, Francisco & Savulescu, Julian (eds.) (2021): *Más (que) humanos: biotecnología, inteligencia artificial y ética de la mejora*, Tecnos, Madrid.
- Larson, Erik J. (2022 [2021]): *El mito de la Inteligencia Artificial*, Shackleton Books, España.
- Lasaga Medina, José (2003): *José Ortega y Gasset (1883-1955). Vida y filosofía*, Biblioteca Nueva, Madrid.



- Latour, Bruno (1996): «Social theory and the study of computerized work sites», en W. J. Orlinokowski & al. (eds.), *Information Technology and Changes in Organizational Work*, Chapman and Hall, Londres, 295-307.
- Lazer, David & al. (2014): «The Parable of Google Flu: Traps in Big Data Analysis», *Science*, 343, 1203-1205.
- LeCun, Yann, Bengio, Yoshua & Hinton, Geoffrey (2015): «Deep Learning», *Nature*, 521, 436-444.
- Llano Alonso, Fernando H. (dir.) (2022): *Inteligencia artificial y filosofía del derecho*, Laborum, España.
- López de Mántaras, Ramón (2016): «La inteligencia artificial y las artes. Hacia una creatividad computacional», en *El próximo paso. La vida exponencial*, BBVA, Madrid, 99-123.
- (2018): «El futuro de la IA: hacia inteligencias artificiales realmente inteligentes», en *¿Hacia una nueva Ilustración? Una década trascendente*, BBVA, Madrid, 160-174.
- (2023): «Inteligencia artificial: grandes retos», conferencia en la Fundación Ramón Areces, 14 de febrero.
- López de Mántaras, Ramón & Meseguer González, Pedro (2017): *Inteligencia artificial*, CSIC, Madrid.
- López Farré, Antonio (coord.) (2020): *La inteligencia artificial en la medicina del tercer milenio. De la predicción al diagnóstico*, Guillermo Escolar, Madrid.
- López Moratalla, Natalia (2017): *Inteligencia artificial. ¿Conciencia artificial?*, Digital Reasons, Madrid.
- Lucas, John R. (1961): «Minds, Machines, and Gödel», *Philosophy*, 36, 112-127.
- Luppardini, Rocci & Adell, Rebecca (eds.) (2009): *Handbook of Research on Technoethics*, Volumen I, IGI Global, Hershey.
- Madrid Casado, Carlos M. (2013): *Hilbert. Las bases de la matemática*, RBA, Barcelona.
- (2015): «Estadística, eugenesia y fundamentalismo científico», *El Basilisco*, 45, 5-31.
- (2017a): *Fisher: La estadística, entre la matemática y la experiencia*, RBA, Barcelona.
- (2017b): *Brouwer. Un geómetra entre la topología y la filosofía*, RBA, Barcelona.
- (2018a): *Filosofía de la Cosmología. Hombres, teoremas y leyes naturales*, Pentalfa, Oviedo.
- (2018b): «¿Qué son las matemáticas? La respuesta de la teoría del cierre categorial», *Berceo*, 175, 163-184.
- (2020): «Filosofía de la Ciencia del Cambio Climático: modelos, problemas e incertidumbres», *Revista Colombiana de Filosofía de la Ciencia*, 41, 201-234.
- (2022a): «Materialism, Logic, and Mathematics», en G. E. Romero & al. (eds.), *Contemporary Materialism: Its Ontology and Epistemology*, Springer, Suiza, 269-301.
- (2022b): «Filosofía de la Medicina: Gnoseología, Ontología y Bioética», intervención en las jornadas *Tecnologías, Ciencias y Ars Medica*, Facultad de Medicina, Universidad de Valladolid, 8 de octubre.

- (2023a): «Filosofía de la Inteligencia Artificial», Lección en la Escuela de Filosofía de Oviedo, 13 de marzo.
- (2023b): «¿Qué es la Inteligencia Artificial realmente existente?», intervención en la jornada *Transhumanismo, Inteligencia Artificial e Inteligencia Emocional*, Instituto Leonés de Cultura, 11 de junio.
- (2023c): «Transhumanismo, inteligencia artificial y biotecnología», intervención de clausura de los 28 Encuentros de Filosofía *Filosofía, Inteligencia Artificial y Nuevas Tecnologías* de la Fundación Gustavo Bueno, 1 de octubre.
- (2023d): «Ética, moral y política de la Inteligencia Artificial», intervención en *E.L.A.I.: Lecciones Éticas de la Inteligencia Artificial*, Fundación Tatiana y Universidad Carlos III de Madrid, 4 de diciembre.
- Marcos, Alfredo (2018): «Bases filosóficas para una crítica al transhumanismo», *ArtefCToS*, 7/2, 107-125.
- (2020): «Información e inteligencia artificial», *Apeiron*, 12, 73-82.
- (2023): «Una inteligencia no tan artificial», *Alfa & Omega*, 22 de junio.
- Marcus, Gary & Davis, Ernest (2014): «Eight (No, Nine!) Problems with Big Data», *The New York Times*, 6 de abril.
- (2019): *Rebooting AI*, Knopf Doubleday Publishing Group, EE.UU.
- (2020): «GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about», *MIT Technology Review*, 22 de agosto.
- Martí, Jordi (2022): «¿Aprender en papel o mediante medios digitales?», *XarxaTIC*, 30 de diciembre.
- Martín Arias, Luis (2023): «¿Pueden ser sustituidos los médicos por máquinas?», intervención en los 28 Encuentros de Filosofía *Filosofía, Inteligencia Artificial y Nuevas Tecnologías* de la Fundación Gustavo Bueno, 29 de septiembre.
- Martín Jiménez, Luis Carlos (2018): *Filosofía de la técnica y de la tecnología*, Pentalfa, Oviedo.
- (2021): *La esencia del Derecho*, Pentalfa, Oviedo.
- McCarthy, John (1979): «Ascribing mental qualities to machines», informe para el Laboratorio de Inteligencia Artificial de Stanford, marzo.
- (2006): «What has AI in Common with Philosophy?», texto corregido de la comunicación presentada en el 14 *International Joint Congress on AI*, celebrado en 1995 en Montreal, Canadá.
- McCarthy, John, Minsky, Marvin L., Rochester, Nathaniel & Shannon, Claude E. (1955): «A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence» [véase el anexo I].
- McGinn, Colin (1993): «Apes, Humans, Aliens, Vampires and Robots», en Paola Cavalieri & Peter Singer (eds.), *The Great Ape Project*, St. Martin's Griffin, Nueva York, 146-151.
- Meléndez Sánchez, Juan (2023): «ChatGPT opina sobre De Tales a Newton», *De Tales a Newton*, 27 de febrero.
- Minsky, Marvin (2010 [2006]): *La máquina de las emociones. Sentido común, inteligencia artificial y el futuro de la mente humana*, Debate, Barcelona.

- Mitchell, Melanie (2019): *Artificial Intelligence. A Guide for Thinking Humans*, Farrar, Strauss and Giroux, EE.UU.
- Moravec, Hans P. (1999): *Robot: Mere Machine to Transcendent Mind*, Oxford UP, Nueva York.
- Moreno, Beatriz & al. (2015): «Eugenesia, la moda del cambio de siglo», en *Historia ilustrada de la Bioética*, ADEMÁS, España, 51-64.
- Müller, Vincent C. (2020): «Ethics of Artificial Intelligence and Robotics», *The Stanford Encyclopedia of Philosophy*.
- Newell, Allen & Simon, Herbert A. (1976): «Computer Science as Empirical Inquiry: Symbols and Search», *Communications of the ACM*, 19/3, 113-126.
- Noah Harari, Yuval (2019a [2015]): *Homo Deus*, Penguin, España.
- (2019b [2018]): *21 lecciones para el siglo XXI*, Penguin, España.
- O’Neil, Cathy (2017 [2016]): *Armas de destrucción matemática*, Capitán Swing, Madrid.
- Obermeyer, Ziad & Emanuel, Ezekiel J. (2016): «Predicting the Future – Big Data, Machine Learning, and Clinical Medicine», *N. Engl. J. Med.*, 375(13), 1216-1219.
- Ongay, Íñigo (2007): «El Proyecto Gran Simio desde el materialismo filosófico», *El Catoblepas*, 64, 1.
- (2024): «Realidad +, + y +. Sobre la filosofía de la realidad virtual de David Chalmers», *El Basilisco*, 59, en prensa.
- Pearl, Judea (2009): *Causality: Models, Reasoning, and Inference*, Cambridge UP, Cambridge.
- (2022): «Los robots hablarán entre ellos, tendrán voluntad propia, deseos... No sé qué le sorprende de esto», entrevista en *XL Semanal*, 23 de abril.
- Pearl, Judea & Mackenzie, Dana (2020 [2018]): *El libro del porqué. La nueva ciencia de la causa y el efecto*, Pasado y Presente, España.
- Penrose, Roger (1989): *The emperor’s new mind*, Oxford UP, Oxford.
- Pérez Álvarez, Marino (2023): *El individuo flotante. La muchedumbre solitaria en los tiempos de las redes sociales*, Deusto, Barcelona.
- Peters, Jonas & al. (2017): *Elements of Causal Inference. Foundations and Learning Algorithms*, MIT Press, EE.UU.
- Picard, Rosalind W. (1997): *Affective computing*, MIT Press, EE.UU.
- Pinker, Steven (1997): *How The Mind Works*, Penguin, Nueva York.
- Richardson, Kathleen (2016): «Sex Robot Matters: Slavery, the Prostituted, and the Rights of Machines», *IEEE Technology and Society Magazine*, 35/2, 46-53.
- Ríos Insua, David & Gómez-Ullate Oteiza, David (2019): *Big data. Conceptos, tecnologías y aplicaciones*, CSIC, Madrid.
- Rodríguez Tauste, Mario (2022): «La mejora (bio)tecnológica y sus enemigos», *El Catoblepas*, 200, 9.
- Russell, Stuart (2019): *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, EE.UU.
- Russell, Stuart & Norvig, Peter (2004 [2003]): *Inteligencia Artificial. Un enfoque moderno*, 2ª edición, Pearson, México.

- (2010): *Artificial Intelligence: A Modern Approach*, 3ª edición, Prentice Hall, Nueva Jersey.
- Ryan, Mark (2020): «In AI We Trust: Ethics, Artificial Intelligence, and Reliability», *Science and Engineering Ethics*, publicado en línea el 10 de junio.
- Sánchez Caro, Javier & Abellán García Sánchez, Fernando (coord.) (2021): *Inteligencia artificial en el campo de la Salud*, Fundación Merck Salud, España.
- Savulescu, Julian (2001): «Procreative Beneficence», *Bioethics*, 15, 413-426.
- (2002): «Deaf lesbians, ‘designer disability’ and the future of medicine», *British Medical Journal*, 324, 771-775.
- (2012): *¿Decisiones peligrosas? Una bioética desafiante*, Tecnos, Madrid.
- Savulescu, Julian & Bostrom, Nick (ed.) (2009): *Human Enhancement*, Oxford UP, Oxford.
- Searle, John R. (1980): «Minds, Brains, and Programs», *Behavioral and Brain Sciences*, 3, 417-457 (incluye la discusión del artículo con otros autores).
- (1985 [1984]): *Mentes, cerebros y ciencia*, Cátedra, Madrid (contiene el capítulo 2 titulado «¿Pueden las máquinas pensar?»).
- (2014): «What Your Computer Can’t Know», *New York Review of Books*, 9 de octubre.
- Singer, Peter (ed.) (2009): *The Cambridge Textbook of Bioethics*, Cambridge UP, Cambridge.
- Skinner, B. F. (1974): *About Behaviorism*, Alfred A. Knopf, Nueva York.
- Slater, P. J. B. (2000 [1998]): *El comportamiento animal*, Cambridge, Madrid.
- Stein-Perlmán, Zach & al. (2022): «2022 Expert Survey on Progress in AI», *AI Impacts*, 3 de agosto.
- Tedre, Matti (2011): «Computing as a Science: A Survey of Competing Viewpoints», *Minds and Machines*, 21/3, 361-387.
- Tegmark, Max (2017): *Life 3.0: Being Human in the Age of Artificial Intelligence*, Knopf, EE.UU.
- (2023): «The ‘Don’t Look Up’ Thinking That Could Doom Us With AI», *Time*, 25 de abril.
- Tomasik, Brian (2014): «Do Artificial Reinforcement-Learning Agents Matter Morally?», *Arxiv*, 30 de octubre.
- Torra, Vicenç (2010): *Del ábaco a la revolución digital. Algoritmos y computación*, RBA, España.
- Turing, Alan M. (1936): «On Computable Numbers, with an Application to the Entscheidungsproblem», *Proceedings of the London Mathematical Society*, s2-42/1, 230-265.
- (1950): «Computing Machinery and Intelligence», *Mind*, LIX, 433-460 (traducido como *¿Puede pensar una máquina?*, KRK, Oviedo 2012).
- Vélez, Iván (2019): *Nuevas mentirosas*, Encuentro, Madrid.
- Véliz, Carissa (2021): «Moral zombies: why algorithms are not moral agents», *AI & Society*, 36, 487-487.
- Vigen, Tyler (2023): «Spurious Correlations», página web consultada el 8 de junio.

- Von Neumann, John (1980 [1958]): *El ordenador y el cerebro*, Antoni Bosch, Barcelona.
- Wang, Pei (2019): «On Defining Artificial Intelligence», *Journal of Artificial General Intelligence*, 10/2, 1-37.
- Warwick, Kevin (2012): «El futuro de la inteligencia artificial y la cibernética», en *Hay futuro. Visiones para un mundo mejor*, BBVA, España, 139-160.
- Weizenbaum, Joseph (1977 [1976]): *La frontera entre el ordenador y la mente*, Pirámide, Madrid.
- Wolfram, Stephen (2023): *What is ChatGPT Doing... and Why Does It Work?*, Wolfram Research, EE.UU.

*Procedencia de las ilustraciones:* Las dos imágenes de la Coda proceden de archive.org y de Wikicommons.



# Índice

<i>Prefacio: El mito de la inteligencia artificial</i> .....	9
<i>PARTE I. GNOSEOLOGÍA DE LA INTELIGENCIA ARTIFICIAL</i> .....	17
1. <i>¿Cuál es el estatuto gnoseológico de la IA?</i> .....	19
1.1. <i>Breve historia de la IA</i> .....	19
1.2. <i>Ideas de técnica, ciencia y tecnología desde la teoría del cierre categorial</i> .....	26
1.3. <i>La IA no es una ciencia sino una tecnología</i> .....	33
<i>PARTE II. ONTOLOGÍA DE LA INTELIGENCIA ARTIFICIAL</i> .....	39
2. <i>¿De qué hablamos cuando hablamos de «inteligencia artificial»?</i> .....	41
2.1. <i>Filosofía espontánea de los profesionales del campo de la IA</i> .....	41
2.2. <i>«La inteligencia humana puede ser reproducida por una máquina»</i> .....	44
2.3. <i>IA fuerte/débil e IA general/específica</i> .....	48
3. <i>Argumentos en contra de la IA</i> .....	51
3.1. <i>Críticos metafísicos de la IA</i> .....	51
3.2. <i>Inteligencia artificial vs. inteligencia emocional</i> .....	53
3.3. <i>Argumento matemático de Lucas y Penrose</i> .....	54

3.4. Argumento de la habitación china de Searle.....	57
3.5. ¿Es ChatGPT realmente inteligente? .....	63
3.6. Argumento de la informalidad de Dreyfus y Larson.....	70
a. Sobreajuste u overfitting.....	73
b. Correlación no implica causalidad ( <i>cum hoc ergo propter hoc</i> ) .....	75
c. IA y medicina.....	78
3.7. La IA realmente existente.....	85
4. La IA no es inteligencia ni es artificial .....	87
4.1. El mito de la inteligencia .....	87
4.2. El mito de lo artificial .....	96
PARTE III. ÉTICA DE LA INTELIGENCIA ARTIFICIAL.....	99
5. Ética, moral y política de la IA .....	101
5.1. Tecnoética y bioética.....	102
5.2. ¿Son los algoritmos o los robots sujetos éticos?.....	110
5.3. Dilemas éticos, morales y políticos de la IA .....	123
a. Invasión de la privacidad .....	123
b. Sesgo en los datos y en los algoritmos.....	125
c. Explicabilidad de las cajas negras.....	129
d. Implicaciones sociales (nuevas mentirosas, desempleo, &c.).....	132
5.4. Indefinición de los marcos normativos vigentes.....	134
5.5. Geopolítica de la IA.....	141
Conclusión.....	149
Coda: Transhumanismo, inteligencia artificial y biotecnología.....	153
Anexo I. Propuesta para el proyecto de investigación de Dartmouth (1955).....	179
Anexo II. «Pausar los experimentos gigantes de IA: una carta abierta» (2023).....	195
Anexo III. Principios de IA de Asilomar (2017) .....	199
Anexo IV. Declaración transhumanista (2009) .....	203
REFERENCIAS BIBLIOGRÁFICAS .....	205



